European Commission

Joint Research Centre

The Role of Artificial Intelligence in Scientific Research

A Science for Policy, European Perspective

Purificato, E., Bili, D., Jungnickel, R., Ruiz Serra, V., Fabiani, J.,
Abendroth Dias, K., Fernandez Llorca, D., Gomez, E.

2025

This document is a publication by the Joint Research Centre (JRC), the European Commission's science and knowledge service. It aims to provide evidence-based scientific support to the European policymaking process. The contents of this publication do not necessarily reflect the position or opinion of the European Commission. Neither the European Commission nor any person acting on behalf of the Commission is responsible for the use that might be made of this publication. For information on the methodology and quality underlying the data used in this publication for which the source is neither Eurostat nor other Commission services, users should contact the referenced source. The designations employed and the presentation of material on the maps do not imply the expression of any opinion whatsoever on the part of the European Union concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries.

The Joint Research Centre: EU Science Hub
https://joint-research-centre.ec.europa.eu

JRC143482
EUR 40448

Print	ISBN 978-92-68-31540-8	ISSN 1018-5593	doi:10.2760/3050242	KJ-01-25-455-EN-C
PDF	ISBN 978-92-68-31427-2	ISSN 1831-9424	doi:10.2760/7217497	KJ-01-25-455-EN-N

Luxembourg: Publications Office of the European Union, 2025

© European Union, 2025

Some content was created using GPT@JRC

The reuse policy of the European Commission documents is implemented by the Commission Decision 2011/833/EU of 12 December 2011 on the reuse of Commission documents (OJ L 330, 14.12.2011, p. 39). Unless otherwise noted, the reuse of this document is authorised under the Creative Commons Attribution 4.0 International (CC BY 4.0) licence (https://creativecommons.org/licenses/by/4.0/). This means that reuse is allowed provided appropriate credit is given and any changes are indicated.

For any use or reproduction of photos or other material that is not owned by the European Union permission must be sought directly from the copyright holders.

Cover page illustration, © DDOK / stock.adobe.com

How to cite this report: Purificato, E., Bili, D., Jungnickel, R., Ruiz Serra, V., Fabiani, J. et al., *The Role of Artificial Intelligence in Scientific Research - A Science for Policy, European Perspective*, Publications Office of the European Union, Luxembourg, 2025, https://data.europa.eu/doi/10.2760/7217497, JRC143482.

Layout: Carmen Capote de la Calle

The Role of Artificial Intelligence in Scientific Research

A Science for Policy, European Perspective

Purificato, E., Bili, D., Jungnickel, R., Ruiz Serra, V., Fabiani, J.,
Abendroth Dias, K., Fernandez Llorca, D., Gomez, E.

2025

CONTENTS

ABSTRACT

Artificial Intelligence (AI) is fundamentally transforming the scientific process across all stages, from hypothesis generation and experimental design to data analysis, peer review and dissemination of results. In many research fields, such as the examined protein structure prediction, materials discovery and computational humanities, AI accelerates discovery, fosters interdisciplinary collaboration and enhances reproducibility, while improving access to advanced analytical and computational capabilities. These developments align with the European Union (EU)'s vision to make AI tools and infrastructure more accessible, strengthening research in areas of strategic importance such as climate change, health, and clean technologies. However, this progress introduces new challenges, including concerns about algorithmic bias, the proliferation of hallucinations and fabricated data, and the potential erosion of critical thinking skills. AI adoption remains uneven across scientific domains, and addressing these risks requires robust governance, transparency and alignment with open-science principles. This report serves as the scientific evidence base for the European Strategy for AI in Science, offering insights to help policymakers navigate the challenges and opportunities of AI. It supports efforts to maximize the benefits of AI for research excellence and competitiveness in the EU, while maintaining a firm commitment to ethical, inclusive, and European values.

FOREWORD

Artificial Intelligence (AI) is transforming today's scientific process, which is the engine driving human progress, in ways that could redefine who participates in, benefits from and has influence over discovery. From the initial stage of inquiry to the publication of results, AI is emerging as a vital tool in **scientific discovery**. This transformative potential, powered by sophisticated models and unprecedented computational capabilities, is expanding access to knowledge, lowering entry-barriers and accelerating breakthroughs from climate to health and clean technologies. Evidence-based policymaking is needed to navigate the paradigm shift underway, which carries major economic, societal and geopolitical implications.

The **AI in Science report** provides the scientific and technical foundation for the **European Strategy for AI in Science**, which aims to define policy priorities for the European Union (EU). Its main objective is to help policymakers maximise AI's benefits for EU research excellence, innovation and competitiveness, while ensuring its deployment remains ethical, inclusive and aligned with core European values.

Drawing on deep dives into AI technology for protein structure prediction, material discovery and computational humanities, this report showcases how AI is accelerating innovation and strengthening EU research. Opportunities are emerging across the full scientific process, from data analysis to the generation of novel research hypotheses. By assessing their potential and impact, the report gathers invaluable insights to guide investments in crucial areas such as high-performance computing and open-science infrastructure. This research can thereby boost **EU competitiveness** at global level.

As AI becomes increasingly integrated in scientific research, it is important to stay mindful of the challenges we may face. The report brings significant concerns to light, such as algorithmic bias, the risk of "hallucinations" and the potential for AI to unintentionally limit the range of research questions being explored. To foster scientific integrity and maintain public trust, we must approach AI in science with a deep commitment to robust governance and a collective vision that emphasises transparency and inclusivity in our scientific practices.

We welcome this report as a timely and crucial contribution to our understanding of AI in science. It shows the Commission's commitment to knowledge-based policy-making. This document is a key resource for a coordinated policy approach as well as a call to collaboration across the scientific community, through shared infrastructures, open-source AI and transparent standards. These efforts would make AI-generated results more trustworthy and consistent, supporting a successful implementation of AI in scientific developments.

Bernard Magenhann
Director-General
European Commission
Joint Research Centre
(JRC)

Marc Lemaître
Director-General
European Commission
Research and Innovation
(RTD)

ACKNOWLEDGEMENTS

The authors would like to extend their gratitude to David Arranz and Eszter Fay for their essential support, to Endika Bengoetxea, Maria Claudia Bodino, Alberto Domini, Kilian Gross, Wide Hogenhout, Daniela Melandri, Nick Vangheluwe, Elinor Wahal and all the colleagues at DG RTD.E.4 and the EU AI Office (specifically DG CNECT.A.2) for their insightful feedback. Equal appreciation and recognition must be attributed to Elisa Calza, Carmen Capote de la Calle, Mario Ceresa, Stephane Chaudron, Giuditta De Prato, Maria Eriksson, Pedro Frau Amar, Gabriele Leoni, Sandy Manolios, Elena Navajas Cawood, Amalia Muñoz Piñeiro, Mauro Petrillo, Yves Punie, Serena Signorelli, Emilie Sundorph and Tobias Wiesenthal, along with the whole JRC.T.3, JRC.T.1 and JRC.F.7 units, for their invaluable contributions and revisions. Finally, a special thanks to all the external experts who reviewed the different deep dives, especially Despoina Tsiafaki from the Athena Research Centre, who supported this study on the archaeology domain. All the groups of colleagues mentioned above have been listed in alphabetical order.

Authors

Erasmo Purificato – European Commission, Joint Research Centre (JRC)

Danai Bili – European Commission, Joint Research Centre (JRC)[1]

Robert Jungnickel – European Commission, Joint Research Centre (JRC)

Victoria Ruiz-Serra – European Commission, Joint Research Centre (JRC)

Josefina Fabiani – European Commission, Joint Research Centre (JRC)

Kulani Abendroth-Dias – European Commission, Joint Research Centre (JRC)

David Fernández-Llorca – European Commission, Joint Research Centre (JRC)

Emilia Gómez – European Commission, Joint Research Centre (JRC)

Disclaimer

During the preparation of this work, the editors used GPT@JRC in order to support the integration process of the contributions by the authors to the final report, as well as to harmonise the narrative style. After using this tool, the editors and authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

GPT@JRC is a platform offering secure access to a wide variety of pre-trained Large Language Models to assist with written office tasks and support scientific work. GPT@JRC is hosted local at the JRC data centre and is part of a JRC-wide study on the potential applications of this new technology within the European Commission. User prompts and generated responses are not shared with third parties.

Note: The last access date for all web sources referenced in this document, unless specified in the individual citation, is 30 September 2025.

[1] Danai Bili is currently with Johnson & Johnson Medical Technology. She contributed to the report while affiliated with the JRC as a trainee.

EXECUTIVE SUMMARY

This report provides a scientific and technical analysis of the role of Artificial Intelligence (AI) in science and the scientific process, offering evidence-based insights to guide strategic decisions. It explores how a wide range of AI techniques, particularly Machine Learning (ML), Deep Learning (DL) and Generative AI (GenAI), are reshaping every stage of research, from hypothesis generation to result publication and community building.

Policy context

AI is reshaping science at an unprecedented pace, transforming how knowledge is generated, experiments are designed, and results are shared. The European Union (EU) is already addressing the broader governance of AI through initiatives such as the **EU AI Act**, which establishes rules to ensure that AI systems used within the EU are safe and respect fundamental rights and values, along with foster trustworthy AI by setting clear, risk-based rules for AI developers and deployers. In parallel, there is a growing need to facilitate and accelerate the responsible uptake of AI in scientific research. The **European Strategy for AI in Science** responds to this need by aiming to develop and improve access to AI tools and computing infrastructure, attract and develop talent, and strengthen research in strategic areas such as climate change, health, and clean technologies. This report provides evidence on how AI is transforming the scientific process, the opportunities it offers, and the challenges it presents. It supports the implementation of the strategy by informing policymakers about the practical implications of AI for research integrity, innovation and EU's competitiveness.

Key conclusions

The analysis conducted in the report confirms that AI has transformative potential across every stage of the scientific process. These capabilities allow researchers to identify patterns and relationships that would otherwise remain hidden, fostering scientific breakthroughs and expanding interdisciplinary collaboration. However, the impact of AI is not inherently positive; it highly depends on the conditions under which it is deployed and governed. The derived findings point to three main areas where policy actions are needed to ensure AI fulfils its potential while safeguarding the integrity of science.

Firstly, the findings reinforce that **open science principles** (including open data, open models and open infrastructure) are paramount for fostering innovation and ensuring the reproducibility and trustworthiness of AI-driven research. Policy support can expand and sustain these ecosystems, ensuring a fair and sustainable access to the required tools.

Secondly, the report highlights the growing challenge of **computational and data infrastructure**. While AI models are becoming more powerful and versatile, they also require significant resources for training and deployment. This makes investment in High-Performance Computing (HPC), AI Factories and open scientific data repositories essential to secure the EU's position as a leader in AI research.

Third, the integration of AI demands a **new skill set for researchers**. The most impactful research emerges from 'hybrid' roles and teams that combine deep scientific domain knowledge with proficiency in AI and data science methods. Policies should therefore focus on attracting, developing and retaining this interdisciplinary talent to ensure that human expertise remains central to the research process.

Finally, a new problem has been identified: the risk of **epistemic drift**. This phenomenon, driven by the co-creation of knowledge with machines, refers to a fundamental shift in what is considered valid scientific knowledge and how it is produced. This can manifest in two ways: as AI technologies inadvertently reinforces established research paradigms and narrows the diversity of questions explored, or as AI technology fosters a culture where scientific conclusions are separated from their human sources and authors, meaning that knowledge production is detached from human oversight and control . AI can also produce fabricated information or 'hallucinations', which, if undetected, may distort scientific understanding. Addressing these issues requires policies that promote AI literacy, critical thinking and multidisciplinary collaboration to ensure human expertise remains central. By proactively mitigating these risks, the EU can safeguard the integrity of the scientific process and foster public trust in science.

Main findings

AI is reshaping research by becoming a collaborative partner in knowledge production and assisting at every stage of the scientific process. It accelerates literature review and knowledge discovery, enables the generation of innovative hypotheses, supports more efficient and targeted experimental design, and facilitates the processing of complex, multimodal datasets. In particular, our main findings suggest that AI is transforming the scientific process of:

- **Asking questions and formulating hypotheses**: AI is evolving from a passive tool into an 'AI co-scientist', assisting at the early stages of research by accelerating literature analysis and identifying knowledge gaps. This function fosters new 'interdisciplinary collaborations', as AI tools can synthesise information from diverse fields to propose novel and testable hypotheses. Large Language Models (LLMs) and other domain-specific systems, in all the steps of the scientific process, can bridge previously unrelated concepts, requiring human oversight to ensure validity and prevent biases towards well-documented areas of research.

- **Designing and conducting experiments**: AI enables automated experiment design, simulation and optimisation, improving efficiency and allowing for 'self-driving labs'. AI tools can generate executable code for experiments and manage complex instruments in real-time. The Nobel Prize-winning success of *AlphaFold* is a prime example of the transformative power of AI, accurately predicting protein structures and accelerating the testing of hypotheses about biological mechanisms. This allows scientists to explore larger experimental domains, but risks oversimplifying real-world complexity.

- **Collecting and analysing data**: AI processes vast, multimodal datasets to detect patterns beyond human capability, a process that has led to breakthroughs in fields like astronomy and genomics. AI-enhanced tools in archaeology analyse massive datasets from satellite and *LiDAR* imagery to discover new ancient sites. Building on this analytical capability, AI is also being used in materials discovery for a new paradigm called 'inverse material design', where the process of analysing properties from existing data is reversed to computationally generate new materials that match a desired set of properties . These processes might raise concerns about data quality, interpretability and the opacity of 'black-box' models.

- **Interpreting results and drawing conclusions**: AI-driven methods assist in translating data into insights and formal theory but often lack transparency in their causal explanations. AI tools can act as a safeguard to ensure new findings are consistent with existing knowledge, aiding in verification.

- **Publishing and communicating findings**: AI tools are increasingly used for scientific writing, including editing and summarisation, enhancing clarity and breaking down language barriers. Automation supports drafting and visualisation, but raises questions about authorship, originality and trustworthiness.

Related and future Joint Research Centre work

This study was developed by different teams of the Joint Research Centre (JRC), specifically belonging to units F.7, T.1 and T.3 (in alphabetical order), and it was managed and overseen by the JRC.T.3's Human Behaviour and Machine Intelligence (HUMAINT) team, whose work focuses on providing a multi-disciplinary understanding of the impact that AI systems have on human behaviour.

This report deliverable is part of the joint 'AI4Science' project between JRC.T.3, DG RTD.E.4 and the EU AI Office (specifically DG CNECT.A.2), which focuses on two different areas. On the one hand, the project analyses the impact of the AI Act on scientific research and outreach the EU scientific community on this aspect. On the other hand, the project provides scientific and technical support to DG RTD and EU AI Office policies intended to foster the use and uptake of AI in scientific research. Future JRC work related to the presented report will follow the described objectives.

Quick guide

AI refers to machine-based systems that, for a given set of inputs and objectives, infer how to generate outputs like predictions or recommendations. The presented report analyses how these systems are used in scientific research. The methodology is mainly based on a broad review of scientific literature, focusing on the core steps of the scientific process. The report highlights that AI offers huge potential for accelerating discovery but comes with significant uncertainties, particularly concerning algorithmic bias, data quality and the risk of generating inaccurate information. The report begins with an introduction that provides a general overview of the role of AI in science and the report's purpose. The core of the report consists of the study of the AI's impact in the scientific process, which explores the recurring trends, challenges and limitations of using AI at each stage. It then delves into three deep dives, which examine how AI is being used in specific research fields, i.e. protein structure prediction, material discovery and computational archaeology, to illustrate its practical applications. The final chapters discuss the ethical, legal and societal implications of AI in science, including data, models and infrastructure, and conclude with final considerations and recommendations for the responsible uptake of AI in scientific research.

1.

INTRODUCTION

1.1 Motivation and approach

Recent developments in the **Artificial Intelligence** (AI) field are having a strong impact in many sectors and activities, being scientific research impacted in different steps and across disciplines. The rapid advancement of AI capabilities, evolving from traditional Machine Learning (ML) to contemporary Generative AI (GenAI) models, is fundamentally reshaping the way knowledge is created and disseminated. This transformation spans multiple fields, affecting the entire spectrum of scientific research and influencing everything from foundational studies to applied technologies.

The European Commission (EC) has recognised this transformative potential and has integrated the promotion of AI in Science into its general AI strategy, intended to harness AI's vast potential while addressing relevant risks. On the one hand, the **European Strategy for AI in Science**[2] intends to accelerate the adoption of AI by European Union (EU) scientists by creating essential enablers such as improved access to data, computational power and talent. On the other hand, the strategy addresses science-specific AI challenges such as preserving scientific integrity and methodological rigour (European Commission: Directorate General for Research and Innovation, 2023).

This report, authored by the EC's scientific service, the **Joint Research Centre** (JRC), aims to provide scientific and technical evidence to the mentioned strategy. It specifically focuses on the use of AI in scientific research, highlighting aspects that are unique compared to its application in other contexts. The core policy problem addressed by this report focus on navigating the rapid transformation of scientific research driven by AI. It seeks to answer the crucial question of maximising the benefits of AI for EU research excellence, innovation and competitiveness, while ensuring that its deployment remains ethical, inclusive, and aligned with EU values. A significant challenge lies in fostering the adoption of AI techniques, which promise to accelerate discovery, enhance reproducibility and promote interdisciplinary collaboration. At the same time, it is essential to mitigate emerging risks such as algorithmic bias, the spread of fabricated data (often referred to as 'hallucinations') and the potential erosion of critical thinking skills. These challenges are not uniform, however, and vary depending on the specific AI technology employed. For instance, the creation of spurious information is a limitation particularly inherent to Large Language Models (LLMs) and other Generative AI (GenAI) systems. This complex issue demands a nuanced, evidence-based approach due to the uneven adoption of AI across various scientific domains and the pressing need for robust governance.

The scope of this report matches the definition of AI system in the EU AI Act (Regulation (EU) 2024/1689), which is in line with the one of the Organisation for Economic Co-operation and Development (OECD, 2024). The term **AI system** means *a machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations or decisions that can influence physical or virtual environments*. In this respect, the report reviews the use of a varied set of AI techniques, including traditional machine learning approaches. However, this report puts a special emphasis on recent developments on AI, notably linked to Large LLMs or GenAI, as the reports addresses for instance their use to deal with large corpus of scientific literature or to support the writing of scientific publications.

The main objectives of this report are therefore to provide a structured, evidence-based analysis that informs and supports the EU's strategic decisions. To this end, the report:

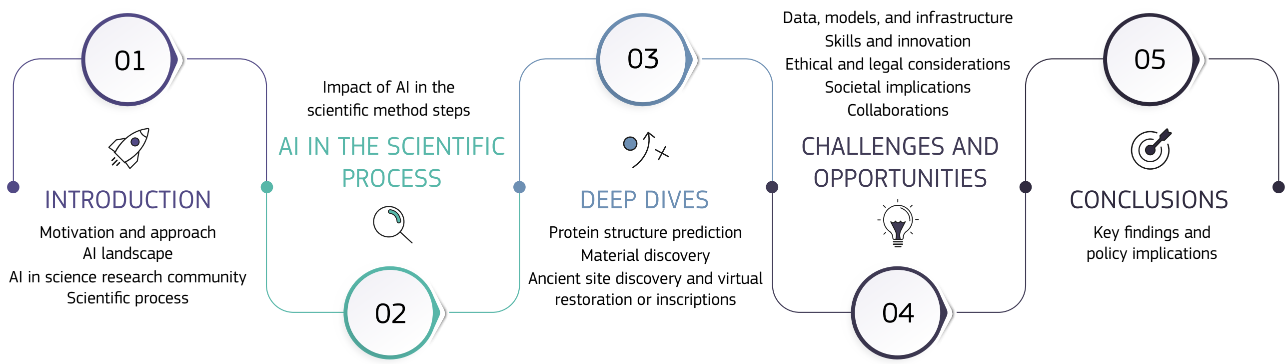
- **Presents an overview** of the current landscape of AI in science and the core stages of the scientific process.

[2] https://research-and-innovation.ec.europa.eu/research-area/industrial-research-and-innovation/artificial-intelligence-ai-science_en.

- **Assesses the impact** of AI across each step of the scientific process, identifying recurring trends, challenges, and limitations.
- **Provides a detailed analysis** through three deep dives, i.e. protein structure prediction, material discovery and computational humanities (specifically ancient site discovery and virtual restoration of inscriptions) to illustrate the practical implications of AI use in diverse fields.
- **Extracts common challenges and opportunities** from these case studies to formulate general conclusions and recommendations for the responsible uptake of AI in scientific research.

This structured approach ensures that the findings can be applied to identify policy problems. For instance, evidence on the computational and data demands of advanced AI models highlights the need for strategic investment in High-Performance Computing (HPC) and open data repositories. Similarly, the detailed analysis of risks, such as fabricated data produced by GenAI models, underscores the need for policies that promote AI literacy, critical thinking and robust governance to safeguard the science integrity. This comprehensive analysis, illustrated by the structure in Figure 1, provides a relevant scientific and evidence-based support to the adoption of the European Strategy for AI in Science.

Figure 1. Visual structure of the report.



Source: JRC's own elaboration.

1.2 Artificial intelligence landscape

This section provides an overview of the overall AI landscape in the EU compared to key global competitors. The JRC's Digital Techno-Economic Ecosystem (DGTES) methodology applied for this analysis of the European and global AI landscape combines and harmonises multiple data sources into a network database representation of the digital ecosystem.[3] In particular, the methodology incorporates a wealth of economic indicators and micro-level information, encompassing industrial, geographical, and technologies dimensions.

The DGTES approach has been deeply developed to map the digital ecosystem, their elements and structure, resulting in the DGTES database (Calza et al., 2022, 2023). The outcome of the mapping exercise leads to a network of players connected through collaborative activities. A *player* is an organisation (academic institution, government body, or firm) that conducts research, innovates or has a business related to digital technologies.[4]

[3] https://joint-research-centre.ec.europa.eu/projects-and-activities/digital-techno-economic-ecosystem-dgtes-mapping-and-analysing-digital-and-other-industrial_en.

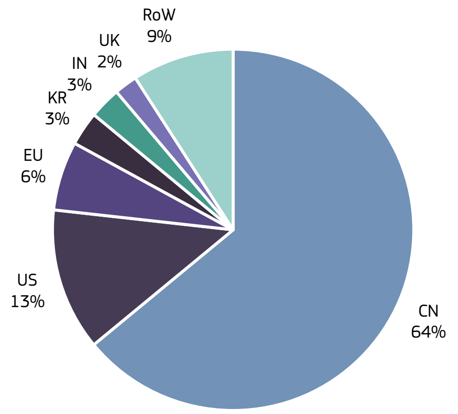
[4] In identifying players, emphasis is placed on organisations rather than individuals, i.e. the applicant or developing

These are the research, (patent-driven) innovation and business *activities* of the digital ecosystem. When more than one player contributes to the same activity, a link emerges between them, resulting in a *network* of collaborations. The analysis employs textual data describing organisational activities globally, identifying 1.9 million entities involved in over 2.5 million relevant activities between 2009 and 2024.[5]

AI is at the core of the digital ecosystem, being the second largest technology out of 15 in the ecosystem in terms of its related activities.[6] These activities are identified by over 700 AI-related keywords entailing various technological solutions based on the principles of and contributing to the development of artificial intelligence, such as Natural Language Processing (NLP) and ML. This section presents analysis on EU's competitiveness on AI, firm ownership, and venture capital investments to assess potential AI foreign funding dependencies in the EU.

The mapping exercise identifies 319 thousand global players engaged in AI during 2009-2024, with over 80% concentrated in China, the US, and the EU (Figure 2). These players are research institutes, government bodies, and firms publishing scientific articles, filing for priority patents, or having their core business related to AI (henceforth, AI activities). Thus, each AI player has at least either its core business oriented to AI, a patent filed on AI, a publication on AI, or any combination of them. On this basis, one player can be linked to multiple activities. On average, AI players have 1.4 activities each.[7]

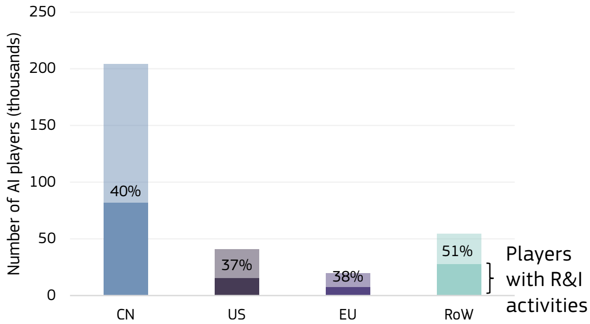
Figure 2. Global distribution of AI players by geographical area, in terms of the cumulative number of firms over 2009-2024.



Source: JRC DGTES database.

By 2024, two in five global AI players had at least one research and innovation (R&I) activity, i.e. had filed for a patent on AI or published a scientific article on the topic. The share is similar across regions, with China displaying a slightly higher share than the US and EU (Figure 3). A closer look by separating research from innovation activities reveals that the EU has a larger composition of AI players with research activities (13% of all AI players) than the US (4%) and China (1%). Taken together, these findings unveil the relevance of innovation activities in the AI global landscape and the role of EU scientific output pushing the AI-knowledge frontier.

Figure 3. Number of AI players with and without R&I activities in the EU, US and China, in terms of the cumulative number of players over 2009-2024 by region. Light-shaded areas represent the number of players with at least one AI-related R&I activity (innovation or research) and labels display their share with respect to all AI players in a geographical area.



Source: JRC DGTES database.

organisation in the context of patents, affiliation instead of individual authors in conference proceedings and publications. Companies' subsidiaries (distinct legal entities) are considered separate players.

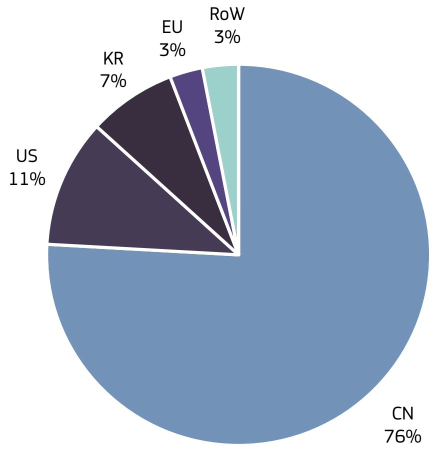
5 Data sources: Scopus, PATSTAT, Orbis, S&P Capital IQ, Crunchbase, Dealroom, and Dow Jones. Data on EU funded projects comes from CORDIS.

6 Digital Areas: 3D Printing; 5G; Advanced Computing, HPC; Artificial Intelligence (AI); Cloud Computing; Cybersecurity; Distributed Ledger Technologies (DLT); Dynamic Data; e-Commerce; Extended Reality; Internet of Things (IoT); Power Electronics; Quantum Technologies; Robotics; Verticals.

7 Data from DGTES database, as on 27 August 2025.

In analysing the **EU's global position in innovation**, the data indicate that the EU contributes a relatively limited portion to the global patent landscape. As displayed in Figure 4, out of more than 226 thousand patents related to AI filed from 2009 to 2024, only 3% were filed by EU players, positioning the EU behind China (76%), US (11%), and South Korea (7%).[8] The disproportionate share of China on innovation activities is consistent with China's patent promotion policies that encourage filing for patents which do not directly imply an improvement in innovation quality (Chen and Zhang, 2019; Long and Wang, 2019).

Figure 4. Number of AI-related patent applications by geographical area, in terms of the cumulative number of innovation activities over 2009-2024 by region.

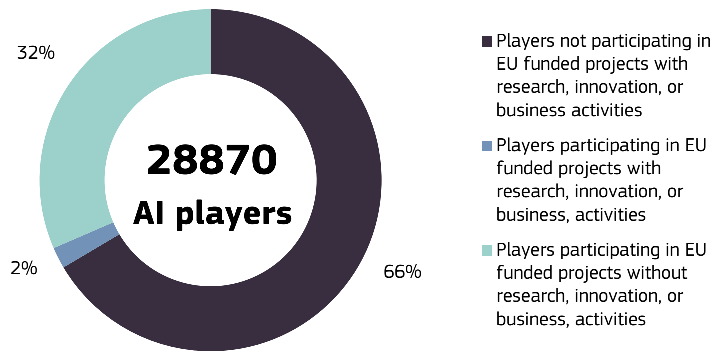


Source: JRC DGTES database.

Funding from the EC plays a crucial role in enhancing EU's capabilities in AI. Because **EU-funded programmes** are accessible only to EU Member States and a select few other countries, data from these programs is excluded from global comparisons for consistency. However, within the EU, programmes like the seventh Framework Programme (FP7), Horizon 2020 (H2020), and Horizon Europe (HE) have been instrumental in strengthening the EU's AI ecosystem and fostering networks of knowledge among participants (Righi et al., 2021).[9]

When considering EU-funded projects, the number of AI players increases by 46%, rising from almost 20000 to 28870, with a third of these players engaged in AI-related EU-funded activities (Figure 5). This underscores the impact of EU-funded projects in supporting the European AI ecosystem, particularly in certain Member States. On average, 36% of AI players in each country are involved in AI-related EU-funded projects (Figure 6). For some countries the relevance of EU programmes for the AI ecosystem is even stronger. For instance, 64% of AI players in Greece participate in EU-funded AI projects. Similarly, Slovenia and Cyprus have more than half of their AI players involved exclusively in such programmes.

Figure 5. Composition of AI-players in the EU by their participation in EU funded projects, in terms of the Cumulative number of AI players in the EU over 2009-2024.

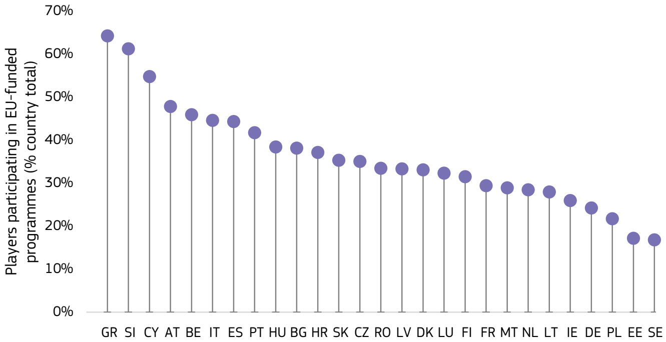


Source: JRC DGTES database.

[8] Differences between US and EU patent share partly reflect institutional rules. The United States Patent and Trademark Office (USPTO) has historically been more permissive toward software while the European Patent Office (EPO) requires a technical character. This leads to more software filings in the US than the EU (OECD, 2009).

[9] Righi et al. (2021) analyses the strategic position of EU countries in AI in R&I network compared to that once accounting for EU funded projects and finds that EC-funded programmes have reinforced the collaborations between players and countries within the EU.

Figure 6. Participation of AI-players in EU funded projects in EU27, in terms of share of AI players in the EU over 2009-2024 by country.



Source: JRC DGTES database.

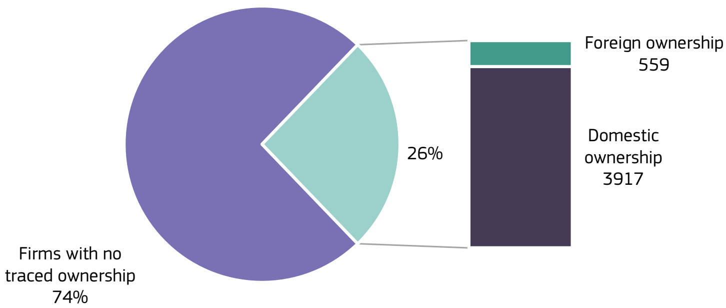
To assess strengths and potential dependencies on **funding of EU in AI**, the focus turns to the investment linkages between AI firms and other entities.[10] In this setting, a firm is financially connected to another firm or entity (which may or may not be an AI-player) through two types of relationships: firm-owner and Venture Capital (VC) investor-investee. While these roles are interrelated, they represent distinct types of connections. A VC investor may influence a firm's strategy without typically having legal control, whereas the firm-owner relationship identifies the entity with formal ownership and control rights, traced through direct and indirect shareholders. VC data is valuable for understanding funding dynamics, while ownership data is essential for mapping corporate control structures.

On the firm-owner link, the complexity of the dataset allows tracing the location of the global ultimate owner of one out of ten firms in the global AI ecosystem. The global ultimate owner is defined as that controlling over 50% of the shares of a company. Foreign-owned firms have a global ultimate owner located in another country. In the EU, a firm is foreign when the owner is located outside the EU, while ownerships between Member States are domestic. With 26% of firms with known ownership, the EU the region with the highest share of tracked ownership (this figure is 9% for China and 15% for the US).

[10] For the analysis of investments, the target are firms, leaving aside other types of players (research institutes and government bodies). The mapping identifies over 17 thousand firms in the EU engaged in AI activities.

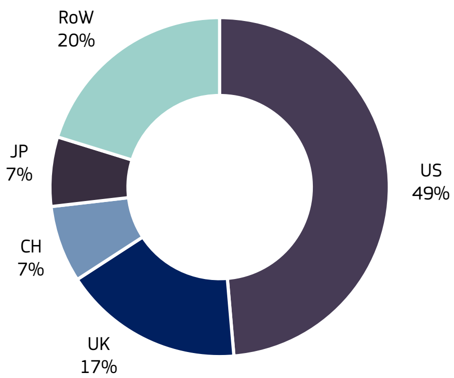
Nine out of ten **EU-based firms on AI** with known ownership data are domestically controlled. This ratio is taken from the over 4.4 thousand EU-based firms on AI with available data on their global ultimate owner (26% of the total) as shown in Figure 7. A closer look at the data, shows that four countries concentrate 80% of global ultimate owners of the over 500 EU foreign-owned firms in AI (Figure 8). US entities control almost half of foreign-owned AI firms located in the EU. The UK hosts 17% of global ultimate owners, while 7% of firms have an owner in Switzerland and another 7% in Japan.

Figure 7. Ownership of EU-based AI firms: data coverage vs foreign composition. Data coverage allows tracing the global ultimate owner of 26% of identified AI firms based in the EU (pie chart). The bar shows the composition of domestic and foreign ownership of the firms with known ownership. Foreign ownership defined as the global ultimate owner holding more than 50% of shares. A firm located in the EU is foreign when the owner is outside the EU.



Source: JRC DGTES database. Data coverage as of 2024.

Figure 8. Foreign ownership of AI firms in the EU by country of the global ultimate owner. Foreign ownership defined as the global ultimate owner holding more than 50% of shares. A firm located in the EU is foreign when the owner is outside the EU.

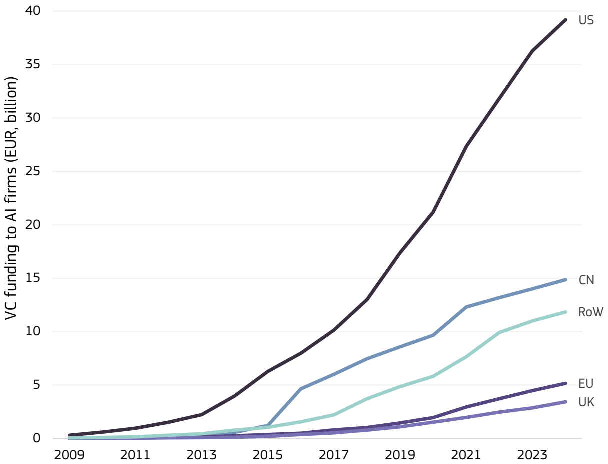


Source: JRC DGTES database. Data coverage as of 2024.

On the **VC investor-investee link**, the final dataset traces back funding received by firms worldwide with a total capital raised of EUR 74 billion during 2009-2024. VC investment is a vital source of resources for innovative startups and small firms with high-growth potential, as they typically have limited access to traditional sources of financing.

AI firms in the EU have limited access to global venture capital funding, despite the various policy actions taken to foster VC investment.[11] Over 2009-2024,[12] 73% of global AI industry VC funding was directed to either US or Chinese firms (53% and 20%, respectively). Only 7% of all VC funding to AI firms during the period went to firms located in the EU (Figure 9), slightly surpassing the funds channelled to the UK based firms (5% of total). Since the outbreak of COVID-19 in 2020, VC investments in AI increased largely leading to a surge in funding for startups. The pandemic accelerated the digital transformation and applications of AI in multiple sectors such as healthcare and mobility and in 2021, global VC investments in AI reached a peak, driven mostly by US based and Chinese firms.

Figure 9. Global VC investments by destination over 2009-2024 (cumulative, EUR billion). The investment figures are weighted by the engagement of each firm in each technology area based on keywords frequency. VC investment rounds include angel, seed, pre-seed, series A- J and unknown types.



Source: JRC DGTES database. VC funding to AI firms as identified by DGTES (Calza et al., 2022).

Figure 10 offers insights into how investment patterns evolve at different phases of business development across main recipient countries, highlighting where investors are focusing their resources within the AI sector over the last five years. Understanding these stages, ranging from early-stage rounds to later-stage investments, can shed light on the maturation process of AI firms and the confidence investors have in their growth potential.[13] These data serve as a critical indicator of the dynamics within the AI startup landscape, illustrating both the scale of investment and the timing and strategic allocations that drive innovation forward.

Compared to international competitors, the US concentrated not only most VC funds, but also

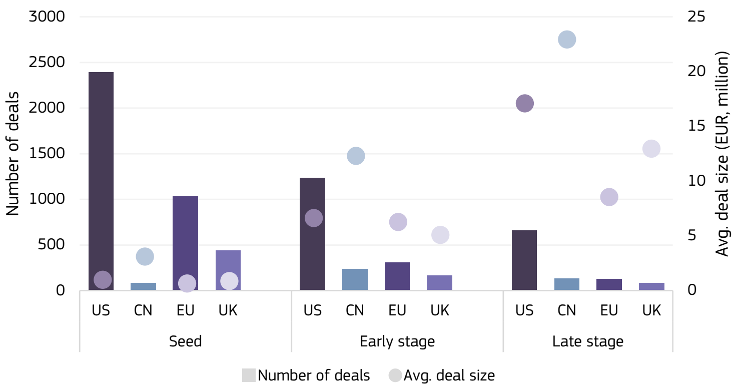
[11] For firms engaged in other technologies besides AI, the total amount of VC funding received is weighted by the relevance of AI for the firm to avoid double counting in the whole digital ecosystem. The weight is then proportional to the number of AI keywords that trigger the appearance of the firm.

[12] Examples of policy actions include the Regulation on European Venture Capital Funds (EUVeCa), the European Fund for Strategic Investment (EFSI), the Pan-European Venture Capital Fund-of-Funds programme (VentureEU), the European Scale-up Action for Risk capital (ESCALAR) programme and InvestEU.

[13] Short description for each stage of deal is as follows – *Seed*: initial funding to develop an idea or prototype, often from angel investors (small round to get a new company off the ground) or seed funds (while the company is young and working to gain traction); *Early*: investment in a startup when it starts scaling its operations, typically Series A or B rounds; *Late*: growth equity firms and more established firms, usually before an initial public offering (IPO) or acquisition. The typical number of seed rounds that a company goes through before completing an IPO is three. However, there is no set number of rounds that must be raised. Data sourced from Crunchbase. For further information on deal types, see https://support.crunchbase.com/hc/en-us/articles/115010458467-Glossary-of-Funding-Types.

the largest number of deals across all stages of funding since 2020. China, however, accounted for the largest deals on average across all funding stages, despite cumulating fewer deals than other regions. Focusing on the EU, the seed stage has been more dynamic than in the UK and even China, adding up to over a thousand registered seed deals of EUR 0.6 million on average. These deals, though smaller on size, provide the initial capital to startups to transform innovative ideas into viable products and services.

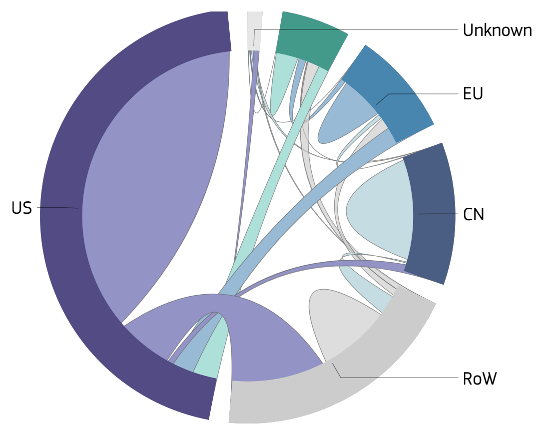
Figure 10. VC deals and size by stage in main destinations over 2020-2024 (EUR million). Bars show the number of deals (y-axis), while dots the average size over the period (x-axis). The investment figures are weighted by the engagement of each firm in each technology area based on keywords frequency. VC investment seed rounds include angel, seed, pre-seed; early-stage rounds include series A and B; late-stage rounds include series C-J and unknown types.



Source: JRC DGTES database. VC funding to AI firms as identified by DGTES (Calza et al., 2022).

Analysing the source of investments and their allocations helps identify potential funding dependencies for innovative firms. The chord diagram in Figure 11 visualises the distribution of VC globally and its source. Since 2020, VC-backed AI firms summed over EUR 28 billion in investments globally. With over EUR 14 billion, the US is the most significant global funding source, with most investment (77%) made locally and the remaining spread across multiple geographical areas. The US has a net outflow with the EU and UK, where US funding for foreign firms is higher than the amount received. China is more restrictive, keeping most investments in the domestic market. With over EUR 2.5 billion invested during the period, the EU occupies the third place as a private funding source and shows a balance between local and foreign investments. There is a net inflow of funding with the UK and the US while a net outflow with China, which represents 5% of all VC-sourced EU investment to AI firms.

Figure 11. Domestic and cross-border VC in main destinations over 2020-2024 (EUR million). Each macro area is represented as a piece of the outer circumference (investor) while the arcs connect to the recipients' macro areas. Arcs that start and end in the same macro area represent domestic investment. Arcs that connect different macro areas, represented the sum of the flows between two selected areas over 2020-2024. The colour of the arc is the colour of the macro area that is attracting more investment between the two. The investment figures are weighted by the engagement of each firm in each technology area based on keywords frequency. VC investment rounds include angel, seed, pre-seed, series A-J and unknown types. The *Unknown* label refers to investors with unidentified location.



Source: JRC DGTES database. VC funding to AI firms as identified by DGTES (Calza et al., 2022).

AI has become one of the defining technologies of the 21st century, shaping competitiveness, productivity, and innovation capacity. Today, the global AI landscape is shaped by three main hubs, namely the US, China, and the EU, with other economies such as the UK, South Korea, and India contributing with the startup and innovation ecosystem. While the US dominates global AI in scale and venture capital investment to fund innovative startups, China has rapidly positioned itself as a global powerhouse in AI patent-driven innovation. The EU, on the other hand, holds strength in scientific output, pushing the knowledge frontier and the region's research excellence. EU funded projects play a pivotal role in boosting the AI ecosystem in the EU, further enhancing its competitive edge.

1.3 AI in science research community

This section provides an overview of scientific papers on **AI in science**, analysing the research community's characteristics, including geographical location, affiliation type and collaboration patterns. Differently from other contributions (European Commission, Directorate General for Research and Innovation et al., 2023; Fudan University and Shanghai Academy of AI for Science, 2025; OECD, 2023) that analyse 'AI in science' by intersecting publications in AI with different scientific fields, this report takes a **more targeted approach**. It specifically examines publications that address topics related to the **scientific process** (as described in Section 1.4) and **metascience**,[14] which studies of how science is conducted, evaluated and disseminated, through the investigation of peer review, reproducibility, research evaluation, research impact, open science and citation analysis (Nature, 2025).

The integration of AI into scientific research represents a broad and fast-evolving interdisciplinary frontier. It spans all stages of the mentioned scientific process, from hypothesis generation and experimental design to data collection, modelling, analysis and interpretation. This transformation is being driven not only by the availability of data and computational infrastructure, but also by the convergence of domain expertise with ML and AI methods.

It is important to distinguish between **multidisciplinary** and **interdisciplinary** approaches. While multidisciplinary research involves drawing on multiple disciplines to examine a topic from various perspectives, interdisciplinary research goes further by integrating the knowledge and methods from these disciplines to create a new, synthesised approach. This distinction is crucial for understanding how AI fosters new collaborative paradigms in science, moving beyond parallel efforts toward a more integrated, problem-focused approach. The resulting research overview reflects a diversity of scientific goals, methodologies and disciplinary intersections, all contributing to the emergence of AI in science.

Practically, the analysis is conducted by adopting **Scopus**[15] as the data source and using the methodology developed by the **divinAI**[16] project, which researches and establishes a set of diversity indicators associated with AI developments, emphasising the geographical patterns of researcher presence in academic and non-academic institutions, with a focus on the differences in their distribution and representation. For a comprehensive description of the methodology, including the definition of the diversity indicators (e.g. affiliation types), as well as recent findings, please refer to the publication by Gomez et al. (2024).

Table 1 illustrates the set of keywords selected for the analysis of the research landscape on AI in science presented in this section (referred to as 'General'), along with the keywords used in the analogous analyses conducted for the deep dives (Section 3). In addition to the reported terms, to tailor the investigation to the intended objectives, the search queries are enriched with AI-related keywords (referred to as *Artificial intelligence*), added in logical conjunction.

14 https://metascience.com/mission/.

15 Scopus is a large, multidisciplinary database of peer-reviewed literature: scientific journals, books, and conference proceedings. Website: https://www.elsevier.com/products/scopus/.

16 https://ai-watch.ec.europa.eu/humaint/divinai_en.

Table 1. Keywords employed for the literature analysis with the divinAI methodology by deep dive.

Topic	Keyword set adopted in divinAI
General	(“discovery” OR “scientific research” OR “enhanced experimentation” OR “hypothesis generation” OR “scientific methodology” OR “scientific methodology” OR “research design” OR “scientific analysis” OR “knowledge discovery” OR “scientific discovery” OR “data interpretation” OR “in science” OR “scientific method” OR “automated scientific discovery” OR “research methodology” OR “scientific process” OR “experimental design” OR “scientific modelling” OR “scientific modelling” OR “scientific inquiry” OR “research innovation” OR “scientific exploration” OR “data-driven science” OR “experimental science” OR “scientific investigation” OR “scientific experiments” OR “scientific research methodologies” OR “scientific research methodology” OR “hypothesis testing” OR “research innovation” OR “research advancement”)
Protein structure prediction	(“protein folding” OR “protein design” OR “protein structure prediction”)
Material discovery	(“materials science” OR “materials engineering” OR “materials chemistry” OR “materials physics” OR “materials technology” OR “nanomaterials” OR “metallurgy” OR “polymer science” OR “ceramic engineering” OR “composite materials” OR “biomaterials” OR “solid state physics” OR “surface science” OR “materials characterization” OR “smart materials” OR “advanced materials” OR “functional materials”)
Site discovery and virtual restoration of inscriptions	(“site discovery” OR “virtual inscription restoration” OR “virtual restoration of inscriptions”)
Artificial Intelligence	(“AI” OR “machine learning” OR “artificial intelligence” OR “deep learning” OR “AI-driven” OR “reinforcement learning” OR “neural networks” OR “predictive analytics” OR “AI technologies” OR “natural language processing” OR “generative ai” OR “large language models” OR “computational intelligence” OR “AI methods” OR “AI methodologies”)

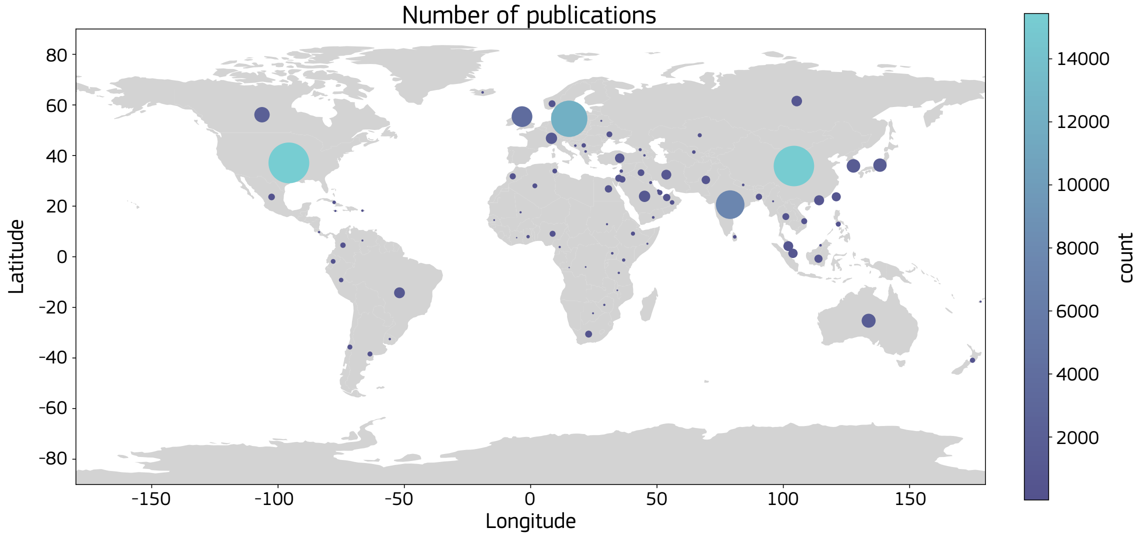
Source: JRC’s own elaboration.

The **general analysis** is conducted on **77,306 publications** indexed in Scopus with the described methodology, considering the contributions having a non-empty affiliation attribute. The selected retrieval period is from **January 2015 and May 2025**. Despite the deep dives (Section 3) being studied in a more recent period (i.e. from 2020 to 2025), this broader analysis considers a comprehensive timeframe to capture the full evolution of the field and the impact of key developments in AI. The year 2015 is indeed considered a crucial turning point due to significant advances in new technologies like DL, notably with the introduction of breakthroughs neural architectures, such as ResNet (He et al., 2016), which laid the foundation for the rapid adoption of AI-related techniques across different scientific disciplines.

This quantitative analysis provides a high-level overview of the geographical distribution of scientific output (Figure 12), the institutional landscape (Figure 13 and Figure 14), and the collaborative networks that characterise the global research effort on AI in science (Figure 15).

In terms of geographical distribution (Figure 12), US and China almost equally lead in total contributions with, respectively, 15,437 (20%) and 15,209 (19.7%) publications, followed by the EU with 12,167 (15.7%) contributions. Further behind, with fewer contributions, are India (7,420; 9.6%), the UK (3,789; 4.9%), Canada (2,020; 2.6%), Australia (1,641; 2.1%), South Korea (1,529; 2%), Japan (1,490; 1.9%), South Africa (1,043; 1.3%) and Switzerland (1,032; 1.3%).

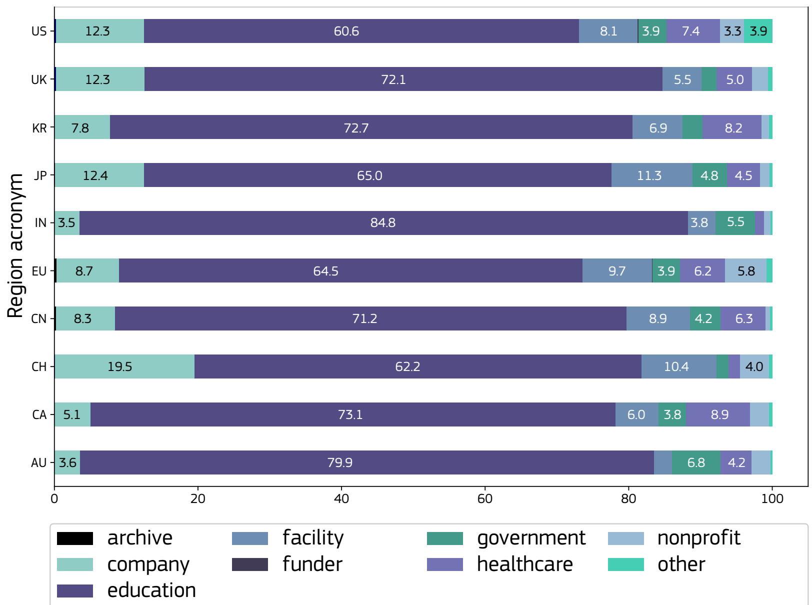
Figure 12. Global geographical distributions of scientific literature contributions related to 'AI in science' topics, indexed by Scopus between Jan. 2015 and May 2025.



Source: JRC's own elaboration based on divinAI methodology.

Conducting a worldwide investigation of the different typologies of establishments involved in the scientific research on AI in science, the analysis of the top ten contributing countries and regions (Figure 13) confirms that the majority of scientific publications are produced by academic institutions. These are followed, in most regions, by private companies, which contribute slightly more than public research facilities and healthcare institutions. Non-profit organisations and government bodies account for a smaller share of output overall. Notably, in India and Australia, government bodies represent the second largest group of contributors after academia, which accounts for 84.8% and 79.9% of all publications, respectively, the two highest academic shares among the top ten regions.

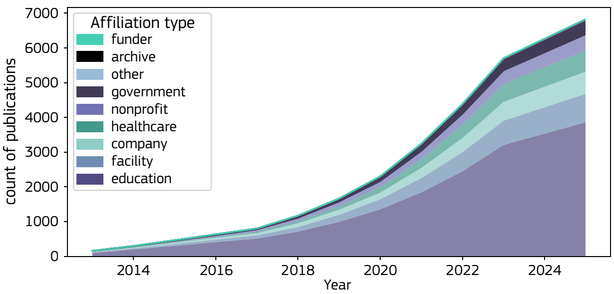
Figure 13. Distribution of 'AI in science' literature contributions by type of establishments and geographical area. Considered contributions are indexed by Scopus between Jan. 2015 and May 2025.



Source: JRC's own elaboration based on divinAI methodology.

A focus on the EU (Figure 14) shows the significant increase in the number of publications over the years, especially since 2020, and confirms the dominant position of educational institutions, with substantial contributions also coming from research facilities, private companies and healthcare institutions, followed by non-profits and government bodies. The growing availability of open datasets, pre-trained models as well as shared computational infrastructure has enabled broader participation across both large and smaller institutions, supporting a more distributed and interdisciplinary research landscape.

Figure 14. Amount of 'AI in science' literature contributions in the EU by type of establishments over the years from Jan. 2015 to May 2025. Considered contributions are indexed by Scopus.



Source: JRC's own elaboration based on divinAI methodology.

Extending the vision on the interplay between different type of institutions and different regions, examining collaboration networks is crucial to understand how the whole research landscape on AI in science is being shaped. Co-authorship analysis (Figure 15(a)) reveals that the most prominent cross-regional collaborations occur between the US and the EU, followed by strong links between the US and China, and to a lesser extent between the US and Canada. Within the Southern Hemisphere, a notable collaboration emerges between Australia and China, while in European continent, the EU and the UK maintain close scientific ties. In the EU, institutional collaborations (Figure 15(b)) among academic institutions, private companies, research facilities, and healthcare institutions are particularly prevalent, with academia playing a central role in facilitating these partnerships. Collaborations between educational institutions and non-profit organisations or government bodies are less common. These patterns reflect the broadening of AI applications across sectors, with academic research increasingly interconnected with clinical, industrial and technological innovation environments.

Figure 15. Collaboration patterns in 'AI in science' field. Amount of co-authored literature contributions by (a) geographical area and (b) type of establishments (only EU). Considered contributions are indexed by Scopus between Jan. 2015 and May 2025.

Source: JRC's own elaboration based on divinAI methodology.

1.4 Scientific process

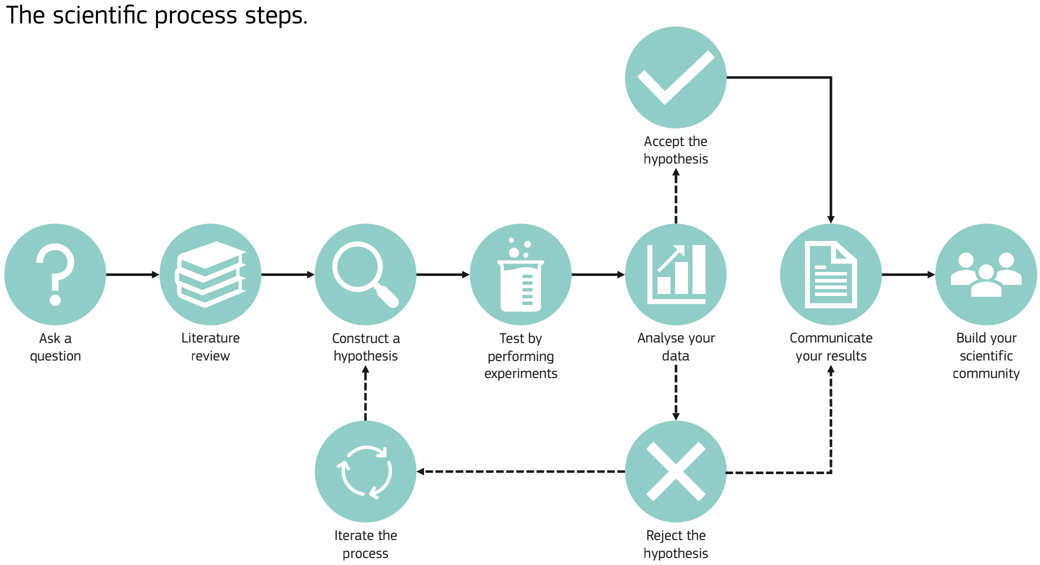
The **scientific process**, often referred to as the *scientific method*, is a structured, iterative framework that sets the ground for the development of empirical knowledge. Originating from the early formulation by the philosopher Francis Bacon, who emphasised inductive reasoning from observation to generalisation (Bacon, 1620), and later refined through Karl Popper's hypothetico-deductive model centred on falsifiability (Popper, 1959), the scientific process has evolved into a widely accepted methodology for systematic investigation. Its core steps (i.e. posing a research question, conducting background research, formulating hypotheses, testing through experimentation, analysing results, and communicating findings) are not rigid rules but guiding principles that support reproducibility and methodological rigor (Dewey, 1910; National Research Council, 2012).

In contemporary science, this process operates as a mechanism of acquiring knowledge, as well as a cognitive and organisational medium that facilitates interdisciplinary collaboration, peer evaluation, and cumulative advancement of theories (Giere, 1979; Hempel, 1966). It serves both epistemic and practical purposes by structuring how questions are asked and answered, allowing for the systematic refinement of hypotheses, and fostering the accountability of empirical claims through standardised procedures.

However, to fully account for how scientific knowledge is validated, institutionalised, and expanded, it is increasingly important to consider the role of the scientific communities as an integral part of the process. They are not merely audiences for scientific findings, but the arena in which credibility, consensus, and quality standards are negotiated (Kuhn, 1962). Building and participating in a scientific community (e.g. through publication venues and institutional collaborations) can be seen as a final and essential step in the scientific process. It is through these communities that knowledge claims are assessed, theories are debated and refined, and interdisciplinary exchanges promote innovation (Knorr-Cetina, 1999; Longino, 2002; Wenger, 1998).

Understanding the impact of AI on each of these stages is crucial to assessing how scientific research is being transformed. From question formulation and hypothesis generation to experimentation, data analysis, dissemination and community engagement, AI technologies are reshaping the way scientists work, the nature of the questions they can pose, and the speed, scale, and precision with which they arrive at and validate new knowledge. Before diving into the details of the novel capabilities and challenges that AI introduces to the core of scientific research, this section will illustrate each step of the scientific process, as displayed in Figure 16.

Figure 16. The scientific process steps.



Source: JRC's own elaboration based on the scientific literature (see Section 1.4).

1. **Ask a question (or make an observation)**. The scientific process begins with observing a phenomenon or posing a question about something of interest, whether broad or specific. This process serves as an empirical approach to assessing phenomena in the universe. The questions typically take the form of *How, What, When, Who, Which, Why* or *Where*. In structured settings, questions should be measurable, ideally with numerical data, to ensure objective analysis.

2. **Conduct background research (or a literature review)**. Conducting background research is a crucial step in the scientific process, as it helps determine what is already known about the topic and whether others have asked the same questions. This research informs the design of an experiment by identifying the most effective techniques and equipment for investigation. Rather than starting from scratch, researchers utilise existing resources to refine their approach and develop a solid theoretical foundation. Understanding the underlying principles behind a study is essential, as it enables scientists to predict outcomes and interpret results accurately. In structured settings, demonstrating this understanding is particularly valuable, as it shows an awareness of why a study yields specific results.

3. **Construct a hypothesis**. A hypothesis is an educated guess that attempts to explain an observation or answer a specific question. It serves as a testable explanation that may later be considered a fact if consistently supported by evidence. A well-formulated hypothesis allows for predictability, ensuring that the observed phenomenon can be tested and repeated under similar conditions. To strengthen its validity, a hypothesis should be accompanied by measurable predictions, enabling systematic evaluation through experimentation.

4. **Test your hypothesis by performing an experiment**. An experiment is designed to test the accuracy of a prediction and determine whether the hypothesis is supported or not. To ensure reliability, the experiment must be a fair test, meaning only one factor is changed at a time while all other conditions remain constant (a condition known as *ceteris paribus*). A well-constructed test should produce measurable or observable changes, allowing for empirical analysis. Repeating the experiment multiple times is essential to confirm that the results are consistent and not due to chance. Controlling for other variables further strengthens the validity of the findings, ensuring that any observed effects are directly linked to the factor being tested.

5. **Analyse your data**. Once the experiment is complete, the collected measurements are analysed to determine whether they support the hypothesis. The results are compared to the initial prediction using the predefined metrics to assess their alignment. A thorough review of all collected data is essential, utilising charts and graphs to identify patterns and trends. This analysis reveals whether the expected outcomes were achieved and provides insights into the findings. By carefully interpreting the data, a clearer understanding emerges of why certain outcomes occurred, offering a logical explanation based on empirical evidence.

6. **Draw conclusions based on acceptance or rejection of the hypothesis**. The conclusion of an experiment summarises whether the results support or contradict the original hypothesis. Key facts from background research can be integrated to provide a clearer explanation of the findings, including any observed relationships between the independent and dependent variables. In many cases, scientists discover that their predictions are inaccurate, leading them to communicate

their results and refine their hypotheses based on new insights. This iterative process is fundamental to the scientific process, as even supported hypotheses often undergo further testing in different conditions. If the results do not align with the original hypothesis, they should be reported transparently rather than altered to fit expectations. Scientists frequently encounter unexpected results, using them as a foundation for formulating new hypotheses and guiding future research. When additional experimentation is necessary, outlining the next steps ensures continuous inquiry and deeper understanding. Scientific research is an ongoing process, and every outcome, whether confirming or refuting a hypothesis, contributes valuable knowledge that fosters further questions and exploration. In structured settings, the emphasis is placed not on proving a hypothesis correct, but on the depth of learning and analytical reasoning demonstrated throughout the process.

7. **Communicate your results**. The final step in a scientific project involves communicating the results through a report, a display board or a presentation. This practice mirrors professional scientific communication, where researchers publish their findings in journals or present them at conferences. Proper documentation of the experiment's results contributes to the broader body of knowledge, allowing other scientists to learn from the findings. Regardless of whether the results support the original hypothesis, they provide valuable insights and may lead to new questions or the formulation of a revised hypothesis for further testing. Scientists value the learning process, and the contributions made through transparent and well-documented research.

8. **Build your scientific community**. The scientific outcomes of individual researchers or teams are mediated by the scientific community they belong to. A scientific community is a network of researchers working on a particular discipline, topic or field, which gathers around a series of publication venues (journals, conferences) or academic organisations. Scientific communities provide the space for peer review, which ensures objectivity and quality of the scientific outcomes, the definition of vocabulary, methodologies, and quality standards and a space for collaborations and debate.[17] Interdisciplinary and cross-institutional activities are also beneficial for scientific progress.

[17] https://en.wikipedia.org/wiki/Scientific_community.

2.

ARTIFICIAL INTELLIGENCE IN THE SCIENTIFIC PROCESS

Artificial Intelligence (AI) is reshaping the way scientific research is conducted, influencing both foundational and practical methods (Jaakkola, 2024; Musslick et al., 2025; Pal, 2023; Rolnik, 2024; Xie et al., 2024). Once confined to computer science and data processing, has become an active participant in scientific inquiry across multiple disciplines like philosophy (Q. Chen et al., 2024; Ye et al., 2024), social sciences (Hou and Huang, 2025; Xu et al., 2024), and medicine (Ahn, 2024; Lin, 2024), especially in the forms of Machine Learning (ML) (Carpenter et al., 2025; Gu and Krenn, 2025a) and Deep Learning (DL) (Onishi, 2025; Wang and Han, 2023), including Generative AI (GenAI) (Alvarez et al., 2024; Goretti et al., 2025) and Large Language Models (LLMs) (Bi et al., 2024; Burton et al., 2024). These technologies are now integrated into various stages of the research process, from generating new ideas to analysing complex data and disseminating findings.

AI tools are transforming how researchers engage with science. They can generate human-like text and images (Bail, 2024; Jeon et al., 2025), automate experimental design (Albert and Billinger, 2025; Bartolomeis et al., 2025), and assist in analysing vast and complex datasets (Ding et al., 2023; Tan et al., 2024). Between 2012 and 2022, the share of scientific publications that engaged with AI across twenty fields quadrupled (Narayanan and Kapoor, 2025). Through 2023, an estimated 1% of academic articles incorporated GenAI during the writing process, with continued growth expected (Gray, 2024). Moreover, AI-assisted papers accounted for 1.57% of selected papers in a 2024 large-scale analysis covering biology, chemistry, geology, materials science, medicine, and physics (Hao et al., 2024). Researchers who integrate AI into their workflows often experience measurable benefits: they publish more, receive citations at a faster rate, and are more likely to take on leadership roles (Hao et al., 2024). These patterns suggest that structural incentives within research ecosystems are accelerating the adoption of AI technologies. The integration of AI offers significant potential. It enables scientists to process large volumes of information quickly, uncover patterns that might otherwise go unnoticed, and automate repetitive tasks (Goretti et al., 2025). These capabilities can enhance efficiency and support more ambitious or creative research (Schmidgall et al., 2025).

However, the integration of AI into science presents significant challenges and risks (Beel et al., 2025; Eger et al., 2025; Hanson et al., 2023; Lahav et al., 2022; Lin and Zhang, 2025; Musslick et al., 2025; Narayanan and Kapoor, 2025; Rolnik, 2024; Schetinger et al., 2023; Tang et al., 2025; Wasim and Zaheer, 2023). There are concerns about issues like hallucination (Drosos et al., 2024; Eger et al., 2025; Galli et al., 2024; L. Huang et al., 2025; Messeri and Crockett, 2024; Schryen et al., 2025), bias (Abeliuk et al., 2025; Algaba et al., 2025; Goretti et al., 2025; Gottweis et al., 2025; Hosseini and Horbach, 2023; Jeon et al., 2025; Kabir et al., 2025; Messeri and Crockett, 2024; Monge Roffarello et al., 2025; Seghier, 2025; Tyser et al., 2024), lack of transparency in how results are produced (Binz et al., 2025; Bolaños et al., 2024; Eger et al., 2025; Gottweis et al., 2025; Saeidnia et al., 2024), and potential to exacerbate social inequality (Bail, 2024; Binz et al., 2025; Goretti et al., 2025). Such limitations threaten the reliability and accountability of AI-assisted science, including a broader risk that widespread adoption may lead to a reduction in diversity in research methods or viewpoints, resulting in a less heterogeneous and understandable scientific landscape (Messeri and Crockett, 2024). Addressing these risks requires careful attention to governance, ethics, and regulatory alignment. The development and use of AI in science should be guided by principles that promote transparency, fairness, and inclusivity. Scientific communities play a central role in determining how these tools are deployed and in managing their benefits and limitations.

This section examines how AI is affecting each stage of the scientific process (depicted in Figure 16), from the initial observation and formulation of research questions to hypothesis development, experimental design, data analysis and communication of results. The analysis draws on peer-reviewed literature and high-impact preprints from January 2020 to July 2025 (with just some few exceptions from previous years), to offer a comprehensive and evidence-based perspective. The section also focuses on methodological transformations and aims to inform researchers, policymakers and interdisciplinary stakeholders who are shaping the future of scientific research. For every scientific process step described previously, patterns and limitations are identified by analysing AI's contributions at each stage .

2.1 Ask a question (or make an observation)

AI can support scientists in the early stages of research by helping them identify gaps, discover new research topics, as well as formulate and refine research questions or objectives (Broska and McFarland, 2025; Feng, 2024; Nicholas et al., 2024). Language models can assist researchers in generating ideas by providing relevant inputs interactively through user-friendly interfaces (Lo et al., 2023; Nigam et al., 2024). Additionally, in this step, AI offers opportunities for defining and framing problems (Pretolesi et al., 2025).

2.1.2 RECURRING AND EMERGING TRENDS

- **Idea generation support** – AI is evolving from a passive 'data-crunching tool' to an active 'idea generator' AI tools act as a *co-scientist* during the early stages of research design and ideation (Feng, 2024; Gottweis et al., 2025). They can go through large amounts of data to find areas where knowledge is limited and suggest new topics (Kim et al., 2024; Nicholas et al., 2024)yet incorporating patient perspectives into health research has been inconsistent. We propose an automated framework leveraging innovative natural language processing (NLP. The AI co-scientist system is designed to generate novel research theories and proposals based on data-driven objectives and guidance (Gottweis et al., 2025). AI can assist in initiating parts of the search strategy for systematic reviews, starting with defining the scope based on the rationale, objectives or questions addressed (Majumder et al., 2024; Wätzold et al., 2024).

- **Knowledge gap identification** – AI is helping to uncover 'what we don't know' by analysing extensive bodies of literature. Modern tools based on LLMs can synthesise diverse pieces of information, making it manageable to identify gaps or inconsistencies in existing knowledge (Oksanen, 2024; Zhang et al., 2024). By creating knowledge graphs or using embedding models, AI systems can conduct a Literature-Based Discovery (LBD), a concept asserting that new knowledge can be uncovered by connecting logically-related fragments of existing information in public literature that have not yet been explicitly linked or interpreted (Luo et al., 2025). Its central idea is to formulate novel hypotheses by bridging previously unrelated concepts found across various publications, as well as hidden relationships rooted in real-world concerns (Z. Chen et al., 2024; Y. Hu et al., 2025). Although this technique was originally developed decades ago (Swanson, 1986), recent advances (such as leveraging DL- or LLM-based analyses on scientific texts) significantly enhance the ability to propose plausible research directions that bridge gaps in the literature (Duede, 2023; Luo et al., 2025). For example, these approaches have been applied to propose drug repurposing candidates and to uncover hidden relations in biomedical text (Gao et al., 2024; Liu et al., 2025).

- **AI as a creative partner** – Researchers have begun folding AI into the creative aspects of science. Instead of using AI only for data analysis, scientists are exploring AI to brainstorm ideas and suggest what questions to pursue (Gu et al., 2025; Pu et al., 2024; Si et al., 2024) providing valuable frameworks for understanding and implementing creative processes. However, recent work using Large Language Models (LLMs. For example, theoretical physicist Mario Krenn developed an AI system (named *Melvin*) that autonomously designed a new quantum optics experiment (Krenn et al., 2016), a setup his team had not conceived on their own. The AI's proposal, initially emailed to Nobel laureate Anton Zeilinger, was novel and intriguing: after several years, the AI-designed experiment worked in practice (Krenn et al., 2022). This success demonstrated that AI could propose viable experiments or hypotheses, effectively asking new scientific questions. Moreover, GenAI models may explore and generate ideas, integrating multidisciplinary perspectives, to solve research problems creatively (Schryen et al., 2025). In this scenario, AI can even generate research topics that experts (e.g. oncologists) rate as novel and significant, reflecting users (e.g. patients) concerns, indicating its capacity to generate ideas in specific interdisciplinary domains (Kim et al., 2024).

2.1.2 CHALLENGES AND LIMITATIONS

AI's strength is often seen in solving pre-specified problems ('easy problem'), while coming up with the problem itself or requiring continual conceptual revision ('hard problem') is still largely beyond current capacities (Battleday and Gershman, 2024). There are concerns that AI might shift collective attention away from new and original questions that lack the data required for AI to demonstrate a benefit (Hao et al., 2024). Automating tasks like identifying a research question can be challenging as they rely on diverse and subjective inputs that may not be structured for machine processing (Musslick et al., 2025). The increasing capability of AI to formulate research questions necessitates a re-evaluation of current scientific paradigms, moving towards 'deep research' agents and potentially AI co-scientists that can operate with significant autonomy (Gottweis et al., 2025). This evolution, however, raises important questions about accountability and the very definition of scientific creativity, prompting a broader discussion on the ethical implications and the evolving role of human scientists, who may increasingly focus on high-level strategic direction, critical evaluation, and the ethical oversight of AI-generated inquiries (Liang et al., 2024).

2.2 Conduct background research (or a literature review)

The rapid growth of scientific literature presents a significant challenge. AI is dramatically improving the efficiency and scope of background research (Bolaños et al., 2024; Mostafapour et al., 2024; Schryen et al., 2025; Jiyao Wang et al., 2024). It helps researchers find and digest information faster (through intelligent search and summarisation), ensures that crucial prior findings are not overlooked (via automated extraction of data and evidence), and even highlights novel connections in literature that can inform new studies (Bednarczyk et al., 2025; Peters and Chin-Yee, 2025; Saeidnia et al., 2024). AI-powered tools are transforming the way researchers handle these tasks by utilising Natural Language Processing (NLP), ML, LLMs, citation and knowledge graphs to automate the retrieval, extraction, and summarisation of scientific information (Glickman and Zhang, 2024; Rajwal et al., 2025; Zeng et al., 2025). They assist researchers in navigating previous work and offer considerable potential for automating literature reviews with personalised models (Agarwal et al., 2025; Z. Liu et al., 2024; McGinness et al., 2024; Miah et al., 2024). Agents based on LLMs have shown the ability to produce readable and detailed literature reviews (Z. Liu et al., 2024; Sami et al., 2024).

2.2.1 RECURRING AND EMERGING TRENDS

– **Enhanced search and discovery** – AI-based tools offer more than just basic keyword matching; they provide contextaware and semantic search capabilities. These tools can generate answers based on search results and offer comparative insights (Eger et al., 2025). Examples include ChatGPT and Gemini 'Deep Research', Elicit, ORKG ASK, NotebookLM, and various Recommender Systems (RSs) (Y. Huang et al., 2025; Oelen et al., 2024; Whitfield and Hofmann, 2023). This is particularly valuable for interdisciplinary work, as AI models, such as LLMs, can efficiently summarise and highlight key findings from vast scientific literature and facilitate the exploration of interdisciplinary research, bridging gaps between different fields (Barman et al., 2025; Eger et al., 2025) . Additionally, graph-based systems help map the relationships between concepts and publications, allowing for the identification of both foundational and emerging works (Gu and Krenn, 2025b; Xu et al., 2025).

– **Automated summarisation and extraction** – AI tools can autonomously summarise abstracts of selected papers, ensuring pertinence to research questions (Sami et al., 2024). They can process diverse unstructured and structured data to uncover hidden patterns and insights within scientific literature (Schryen et al., 2025). Named-entity recognition and classifiers can be used to extract specific entities or concepts from articles (Bolaños et al., 2024). These systems help scientists stay up to date with new publications, identify relevant findings, and quickly gain an overview of a field.

– **Information extraction and organisation** – Beyond summaries, AI is being used to extract structured knowledge from papers automatically (Bernasconi et al., 2022; Dagdelen et al., 2024). An example is the Scite Assistant's AI-powered tool for reading research tables and data (Lund and Shamsi, 2023). Instead of manually examining articles for specific data points or experimental results, researchers can leverage AI to pull out and organise those details (Hsu et al., 2024). This is particularly valuable in fields like pharmacology or clinical research, where missing a critical data point could significantly alter conclusions (Xu et al., 2025). By structuring methods and outcomes from hundreds of papers, AI can enable more efficient meta-analyses and evidence aggregation.

2.2.2 CHALLENGES AND LIMITATIONS

Despite advancements, challenges persist, including data quality and coverage gaps, bias in AI models influencing the visibility of research, and scalability issues (Eger et al., 2025). GenAI's effectiveness depends on the specific knowledge development activity in the review; while helpful for identifying and synthesising literature, it may fall short in critical analysis or aggregating complex evidence (Schryen et al., 2025). Some studies have shown inconsistencies in performance for literature searches across different AI tools (Lund and Shamsi, 2023; Schryen et al., 2025). Relying on AI tools without reading the actual papers can lead to the invention of references or spurious correlations (Buriak et al., 2023), a growing concern for funding agencies and review committees who have reported encountering non-existent publications in submitted bibliographies . Excessive reliance on LLMs for literature synthesis can hinder the development and refinement of conceptual frameworks, which are essential for scientific education.

2.3 Construct a hypothesis

AI can play a crucial role in generating hypotheses, a fundamental step in the scientific discovery process (Abdel-Rehim et al., 2025; Barman et al., 2025; Batista and Ross, 2024; Eger et al., 2025; Liu et al., 2025; Misra and Kim, 2024). It can propose credible connections based on existing literature, predict novel links using ML models, suggest formal theoretical conjectures, and even derive specific potential formulas (Battleday and Gershman, 2024; Beel et al., 2025; Luo et al., 2025). LLMs can generate ideas that are plausibly novel and feasible, comparable to those produced by human researchers (L. Li et al., 2024). Agent-based systems, such as the AI co-scientist, are specifically designed to automate the entire research pipeline, including hypothesis formulation and helping uncover new and original knowledge, as well as to formulate demonstrably novel research hypotheses and proposals (Gottweis et al., 2025).

2.3.1 RECURRING AND EMERGING TRENDS

- **Literature-based hypothesis generation** – AI systems systems can generate research hypotheses by synthesising extensive literature and identifying latent relationships (Gottweis et al., 2025; Saeidnia et al., 2024). Multi-agent systems leveraging knowledge graphs and LLMs, like SciAgents, can autonomously generate and refine hypotheses, even revealing hidden interdisciplinary relationships (Ghafarollahi and Buehler, 2024). AI can be used to design experimental stimuli or interview questions, which are linked to hypothesis formulation (Feng, 2024). In practice, the same AI tools that identify research gaps can also propose specific hypotheses. By analysing patterns in scientific papers, AI might suggest, for example, that a certain protein could be a key regulator in a disease (connecting two previously unrelated study findings) (Cheerkoot-Jalim and Khedo, 2021; Henry and McInnes, 2017).

- **ML-driven link prediction** – In many fields, hypothesis generation can be framed as a link prediction or pattern completion problem (Krenn et al., 2023). These predictions are essentially hypotheses about cause-effect or interactions. A concrete example is in drug discovery: ML models trained on known drug–target interactions have proposed new candidate drug molecules for specific targets (Dara et al., 2022; Sadybekov and Katritch, 2023). Notably, in early 2020, a DL approach identified a novel antibiotic drug (later named 'halicin') by predicting antibacterial activity for molecules in a large chemical library (Stokes et al., 2020). The model generated the hypothesis that halicin would be effective against certain resistant bacteria, a hypothesis later confirmed experimentally, showcasing AI's ability to hypothesise useful new therapeutics.

- **AI in theoretical discovery** – AI has even started contributing hypotheses in abstract domains like mathematics (Jejjala et al., 2019; Lample and Charton, 2019; Raayoni et al., 2021). In late 2021, DeepMind's team combined ML with human insight to conjecture new mathematical theorems (Davies et al., 2021). Their AI system analysed massive datasets of mathematical objects to spot patterns and suggested conjectures in areas such as knot theory and *representation theory*. A relevant finding of the study revealed that AI can aid in formulating hypotheses (*conjectures*) even in fields with no experimental data, by guiding human intuition toward fruitful ideas.

- **Symbolic AI and hypothesis formation** – AI techniques (like *automated reasoning* and *symbolic regression*) are being used to formulate scientific laws or hypotheses (Angelis et al., 2023; Reddy and Shojaee, 2025). Instead of a human guessing a functional form, symbolic regression algorithms search the space of mathematical expressions to fit data and produce human-readable hypotheses (i.e. *equations*) (Makke and Chawla, 2024). For example, researchers have developed AI methods that rediscovered Kepler's third law of planetary motion and Einstein's time-dilation formula by combining logical reasoning with a small amount of data (Cornelio et al., 2023). The past few years have seen improvements in ensuring such AI-generated hypotheses obey known physics and are consistent with known principles (Shin et al., 2024).

2.3.2 CHALLENGES AND LIMITATIONS

AI systems trained on existing literature may favour popular paths, potentially reinforcing established research paradigms and neglecting underrepresented directions (Eger et al., 2025; Tang et al., 2025). AI-generated hypotheses may lack transparency, making it difficult for researchers to verify their scientific soundness or underlying assumptions (Eger et al., 2025). There is a critical concern that a considerable part of AI-generated research documents are plagiarised, bypassing detectors and not acknowledging original sources (Gupta and Pruthi, 2025). Reliance on AI might lead to a homogenisation of ideas across populations (Gottweis et al., 2025). AI seems more inclined to stimulate answers to existing issues rather than generate new ones, potentially slowing the expansion of knowledge by inducing collective hill-climbing (Hao et al., 2024).

2.4 Test your hypothesis by performing an experiment

AI is revolutionising the experimentation process by enhancing the design and execution of research activities (Albert and Billinger, 2025; Bartolomeis et al., 2025; Q. Huang et al., 2024; Z. Liu et al., 2024). By intelligently selecting and automating experiments, AI enables researchers to focus on the most promising avenues of investigation. Its capabilities allow for the operation of experiments with minimal human intervention, providing real-time interpretation of results (Stevenson et al., 2024; Yacoub et al., 2022). This efficiency facilitates quicker testing of hypotheses and allows for the exploration of larger experimental domains than ever before. Moreover, robotics combined with AI-driven automation has simplified laboratory procedures, making it easier for scientists to conduct complex studies (Rolnik, 2024). AI algorithms significantly contribute to the advancement of science by simulating experiments, predicting outcomes, and optimising various conditions, thereby accelerating scientific breakthroughs (Feng, 2024). These tools also provide valuable assistance in programming, enhancing the implementation of statistical analyses (Goretti et al., 2025). Additionally, specialised systems like the ‘agent laboratory’ are designed to function within automated research pipelines, further simplifying the scientific process (Battleday and Gershman, 2024; Beel et al., 2025; Schmidgall et al., 2025; Tom et al., 2024).

2.4.1 RECURRING AND EMERGING TRENDS

- **Experiment design and optimisation** – AI techniques (including *Bayesian optimisation* and *reinforcement learning)* are used to design experiments that yield maximal information (Dworschak et al., 2022; Greenhill et al., 2020; Treloar et al., 2022). Instead of exhaustive or random trial-and-error, scientists can employ ML to prioritise which experiments to run, essentially testing the most promising parts of a hypothesis first (Dang et al., 2024; Mece et al., 2020). In materials science and chemistry, this approach has given rise to ‘self-driving labs’ (Tom et al., 2024). For example, an autonomous chemical lab can be set up as a closed-loop system: an AI model proposes a set of experimental conditions, a robotics system runs the experiment and measures results, and the AI learns from the outcome to propose the next experiment. Recent reports describe autonomous labs that can execute 50–100 times more experiments per day than a human[18] (Biron, 2023). By operating 24/7 and adjusting protocols on the fly, these AI-driven labs significantly accelerate hypothesis testing optimisation

[18] https://newscenter.lbl.gov/2023/04/17/meet-the-autonomous-lab-of-the-future/.

and free human scientists to focus on designing high-level hypotheses (Delgado-Licona and Abolhasani, 2023; Desai et al., 2025). Moreover, AI frameworks support automating tasks like hyperparameter tuning (Czako et al., 2021; Shawki et al., 2021), as well as generating executable code for experiments (Wills et al., 2024; Fengxiang Zhao et al., 2024; Zhuang and Lin, 2024).

- **Simulation and data generation** – AI is fundamentally changing the nature of scientific experiments by enabling large-scale simulations. It can be used to simulate human behaviour for research purposes, such as in survey research, online experiments, and agent-based models (Albert and Billinger, 2025; Bail, 2024; Gürcan, 2024; Yiren Liu et al., 2025; Rosala and Moran, 2024). AI can also generate simulated or synthetic data, although only a few biomedical researchers reported using it for this purpose (Afonja et al., 2025; Goyal and Mahmoud, 2024; Ruediger et al., 2024). The ability to simulate known results with AI may indicate functional abilities in data generation and theory building (Afonja et al., 2025; Christou, 2023; Lehr et al., 2024; R. Li et al., 2024). This is evident in fields like weather forecasting (Conti, 2024), where AI models can process billions of data points more quickly than traditional physics-based models, and in structural biology, where models like AlphaFold accurately predict protein structures (Jumper et al., 2021). AI's ability to analyse patterns from data also allows it to be used to solve a wide range of mathematical problems, from basic algebra to advanced calculus (Davies et al., 2021).

- **Controlled instruments** – In large-scale scientific experiments, such as physics detectors, telescopes, and particle accelerators, AI plays a crucial role in managing complex instruments and identifying noteworthy events (Jiao et al., 2024; Zubatiuk and Isayev, 2021). For instance, particle physics experiments employ ML models in real-time to decide which collision events to record, a form of hypothesis testing that involves retaining data that might confirm a theory (Jiao et al., 2024; Krenn et al., 2022; Lai et al., 2022). In astronomy, AI pipelines analyse telescope data nightly to identify phenomena like supernovae or new asteroids for follow-up, effectively testing hypotheses about *transient events* by catching them as they happen (Fluke and Jacobs, 2020; K. Huang et al., 2024; Kodi Ramanah et al., 2022). These applications show how AI increases the responsiveness of experimental tests.

2.4.2 CHALLENGES AND LIMITATIONS

The speed and volume at which AI can design and execute experiments can lead to insufficient ethical oversight and inadequate safety controls (Eger et al., 2025; Jeon et al., 2025; Seghier, 2025). There are concerns about AI models fabricating data and studies, which can be difficult to ascertain without significant time for review (Kabir et al., 2025; Saeidnia et al., 2024). AI models can also propagate coding errors and biases present in their training data (Becker et al., 2023; Kücking et al., 2024; Straw, 2020). Handling AI-generated hallucinations, such as in citation creation for experimental methods, can result in inaccurate or non-existent references (L. Huang et al., 2025; Monge Roffarello et al., 2025). Regarding the generation of code for experiments, translating methodology into executable actions remains a challenge for AI agents without dedicated interfaces or tools (Becker et al., 2023; Huang et al., 2024).

2.5 Analysing data from experimental results

AI tools are increasingly being utilised to assist in data analysis and interpretation (Bi et al., 2024; Drosos et al., 2024; Eger et al., 2025; X. Hu et al., 2024; Liu et al., 2023; Rolnik, 2024; Wachinger et al., 2024). ML algorithms excel at identifying patterns, correlations, and trends within large datasets (Pal, 2023; Wasim and Zaheer, 2023). This area of application is often referred to as 'AI as Quant' (Weiskopf, 2024). AI can handle complex datasets, which may enhance or even surpass human cognitive abilities in examination (Musslick et al., 2025). Additionally, AI tools can aid in both data cleaning and analysis (Binz et al., 2025; Monge Roffarello et al., 2025; Zhu et al., 2025).

2.5.1 RECURRING AND EMERGING TRENDS

- **Automated data processing** – Modern scientific experiments and observations often produce data on scales that overwhelm traditional analysis (Pagliaro and Sangiorgi, 2023). AI systems can automate data collection and cleaning, ensuring higher accuracy and consistency (Rolnik, 2024; Zhu et al., 2025). They enable efficient data processing, pattern recognition, and knowledge extraction from various sources (Lipkova et al., 2022; Fei Zhao et al., 2024). AI can analyse large volumes of text, audio, video or any data to uncover patterns and trends in cultural, historical, and social phenomena (Rolnik, 2024; The Royal Society, 2024). In quantitative analysis, AI methods like rule-based learning, supervised learning, and LLMs can assist with tasks like inductive coding of qualitative data (Fengxiang Zhao et al., 2024).

- **Pattern recognition** – AI has stepped in as a robust pattern recogniser (Jalaian and Bastian, 2023; Serey et al., 2023). In high-energy physics, for example, billions of collision events are produced at the LHC, far too many for humans to scrutinise individually. Deep Neural Networks (DNNs) are trained to recognise the *fingerprints* of interesting particles within this deluge. An Argonne Lab study used a Neural Network (NN) autoencoder to sift through LHC data and detect subtle anomalies that could indicate new physics beyond the Standard Model (ATLAS Collaboration, 2024). Similarly, in astronomy, DL has been used to classify millions of galaxies and identify new exoplanets. NASA's ExoMiner NN recently validated 301 new exoplanets from Kepler telescope data in one large batch, by learning to distinguish true planetary signals from false positives (Valizadegan et al., 2022). These successes highlight AI's ability to analyse massive datasets efficiently, uncovering patterns (e.g. new planets, potential new particles, etc.) that would have taken prohibitive human effort to find.

- **Advanced data mining and multivariate analysis** – AI enables scientists to analyse data with many variables in ways traditional statistics cannot (Kumar et al., 2023). For instance, in genomics and systems biology, ML models can integrate thousands of features (such as genes, mutations, and expression levels) to identify clusters or predictive patterns related to disease (De La Vega et al., 2021; DeGroat et al., 2023). In climate science, DL models digest petabytes of simulation output and observational data to detect trends or extreme events (Jiang et al., 2022; Salcedo-Sanz et al., 2024). What is crucial is that AI can model highly complex, non-linear relationships in data. This leads to novel discoveries: ML analysis of seismic data has identified previously unknown micro-earthquake patterns (Mousavi and Beroza, 2023; Yang et al., 2021), and DL on network traffic has revealed telltale signs of cosmic events, such as identifying gravitational lensing patterns in telescope images (Kim et al., 2021). AI-driven data analysis typically demonstrates increased sensitivity and uncovers novel insights, as algorithms can detect subtle signals that are often overlooked by traditional analyses.

- **Automated interpretation and understanding** – AI-driven methods increasingly support the interpretation of scientific data by automating complex analytical tasks (Gil et al., 2021; Sarker, 2022). In domains such as medical imaging and microscopy, DL models not only detect relevant features (such as tumours in scans (Gharaibeh et al., 2022; Kao and Yang, 2022)) but also quantify and characterise them, enabling more efficient and precise assessments (X. Li et al., 2024). Similarly, in particle physics, ML algorithms can infer physical properties, such as charge or momentum, from raw detector outputs, often outperforming traditional reconstruction methods (Karagiorgi et al., 2022). In environmental sciences, AI-based surrogate models emulate computationally expensive simulations, such as climate or fluid dynamics models, allowing for real-time analysis of dynamic systems (Emmerich et al., 2024; Jones et al., 2023). Moreover, AI supports the integration of quantitative and qualitative analyses, for instance by semantically enriching social media or experimental data and aligning it with external sources such as literature or code repositories (Bryda and Sadowski, 2024; Rietz and Maedche, 2021).

2.5.2 CHALLENGES AND LIMITATIONS

AI tools are likely to make mistakes or hallucinate in analysis (L. Huang et al., 2025). The use of AI in data analysis, especially for complex or nuanced data, carries risks of illusions of understanding if contextual sensitivity and multidisciplinary perspectives, often preserved by qualitative approaches, are stripped out (Messeri and Crockett, 2024; Weiskopf, 2024). In this scenario, transparency and explainability of AI models are crucial, as complex models can obscure the process by which predictions are made (Chowdhury et al., 2023; Mengaldo, 2025; Walmsley, 2021). Over-reliance on AI analysis could diminish researchers' critical thinking skills (Schemmer et al., 2023; Zhai et al., 2024). Evaluating the reliability and trustworthiness of AI results is crucial, necessitating validation and a critical review by human experts (Tsamados et al., 2025). There are concerns about the potential for over-concentration of AI research leading to redundant innovation rather than novel insights (Doshi and Hauser, 2024; Hao et al., 2024).

2.6 Draw conclusions based on acceptance or rejection of the hypothesis

Once data are analysed, scientists must draw conclusions, as determining whether a hypothesis is supported, and formulate theoretical explanations. AI algorithms can assist researchers in deriving meaningful conclusions based on the patterns and trends identified in the data analysis procedures (Pal, 2023; Wasim and Zaheer, 2023). Frameworks and systems like *AIGS* aim to autonomously complete the entire research process or contribute to theory building (Lehr et al., 2024; Z. Liu et al., 2024). AI helps *connect the dots* between data and theory. It can elevate raw analysis to formal theory, assist in validating that conclusions are causally sound and consistent with known science (Burstein and LaFlair, 2024), and even participate in the reasoning process (Ghafarollahi and Buehler, 2024; X. Liu et al., 2024). The result is that scientific findings can be reached more rigorously and, at times, more creatively, with AI offering a second pair of eyes to catch errors or propose explanations that humans might miss (Krenn et al., 2022; Wang et al., 2023).

2.6.1 RECURRING AND EMERGING TRENDS

- **Falsification and verification** – AI systems can be designed with explicit falsification components to identify and verify potential scientific discoveries based on experimental results (K. Huang et al., 2025; Z. Liu et al., 2024). The process of falsification is considered central to AI-generated science, built on experimentation and aimed at fostering creativity.

- **Causal inference and hypothesis evaluation** – Drawing correct conclusions often hinges on distinguishing correlation from causation. AI is aiding scientists in performing causal inference on complex datasets (Chernozhukov et al., 2024; Wang et al., 2022). Advanced methods can suggest causal relationships that explain the data. For example, AI algorithms have been applied to epidemiological data to infer causal links between risk factors and outcomes, guiding conclusions about disease etiology (Jacquot et al., 2023). In fields like economics or social sciences, where randomised trials are impractical, ML-based causal models help researchers conclude which factors truly have effects (Schölkopf, 2022). Additionally, AI can run counterfactual simulations: using generative models to simulate *what if* scenarios that test the robustness of conclusions under different assumptions (Kirfel et al., 2025).

- **Ensuring consistency and validating conclusions** – AI is also used as a safeguard to check that new findings make sense in the context of existing knowledge (Cheong, 2024). Researchers have proposed formal verification methods where an AI system evaluates whether a learned hypothesis (from an ML model) violates known physical laws or constraints (Reddy and Shojaee, 2025). This alignment with scientific principles is essential as AI takes a larger role in discovery, and tools are being developed to integrate domain knowledge as a constraint on AI-driven conclusions (Gabriel, 2020; Reddy and Shojaee, 2025). Moreover, by analysing statistical patterns and metadata from published studies, AI can estimate a study's probability of being true upon replication and foresee unreliable results (Hong et al., 2023; Ryan, 2020).

- **Human-AI collaborative reasoning** – Drawing conclusions often benefits from explanation and reasoning, where AI can assist via LLMs or knowledge-based systems (Forer and Hope, 2024; Ifargan et al., 2025). LLMs, fine-tuned on scientific knowledge, can assist researchers by logically narrating the implications of the results or by suggesting additional tests to verify a conclusion (Cohen and Peled, 2025; Singhi et al., 2025). When grounded in data and adequately verified, these models can help researchers articulate conclusions more clearly or consider alternative interpretations.

2.6.2 CHALLENGES AND LIMITATIONS

Despite its transformative potential, AI-assisted research presents notable epistemological and methodological limitations. In domains heavily reliant on quantitative or large-scale data, AI can foster a form of data centrism that privileges inductive pattern recognition over theoretical reasoning, weakening the dialectical interplay between empirical evidence and conceptual frameworks (Kumar et al., 2024; Patel, 2024). Moreover, current AI systems often lack the capacity for deep interpretation or contextual judgment, making them adept at reproducing established knowledge but poorly suited for distinguishing between spurious correlations and groundbreaking insights (Buriak et al., 2023; Wang et al., 2023). Paradoxically, the precision afforded by AI tools can obscure rather than clarify scientific understanding, raising concerns about the depth and originality of conclusions derived from such systems. Messeri and Crockett (2024) warn that the proliferation of AI tools in science risks *introducing a phase of scientific enquiry in which we produce more but understand less*, potentially leading to scientific monocultures where certain methods and viewpoints dominate, making science less innovative and more prone to errors. For conclusions to be truly interdisciplinary and robust, human scientists must retain critical responsibilities for reviewing, critiquing, validating theories, identifying gaps, and extending knowledge, acting as directors and overseers of AI-generated insights.

2.7 Communicate results

AI-based systems are influencing how scientists communicate. When used responsibly, these tools can greatly aid researchers in scientific writing, editing, and publishing (Eger et al., 2025; Goretti et al., 2025; Jaakkola, 2024; Melliti, 2024; Ruediger et al., 2024; Zhuang et al., 2025). They are instrumental in streamlining tasks such as manuscript formatting, reference expansion, checking grammar, and enhancing clarity (Feng, 2024; Lin, 2025; H. Wu et al., 2023). AI can assist with language tutoring for non-native English speakers (Pang et al., 2025), as well as in drafting papers, generating specific sections like titles, abstracts, and related work, and assisting with citations (Salvagno et al., 2023; Wallwork, 2024). LLM-based systems are being explored for automating the whole paper-writing process (Liang et al., 2024).

2.7.1 RECURRING AND EMERGING TRENDS

- **AI-assisted scientific writing** – AI tools , particularly LLMs, are increasingly employed as writing assistants throughout the scientific communication process. By supporting researchers in overcoming initial writing barriers, refining grammar, and improving argumentation structure, these systems help streamline manuscript preparation and enhance textual clarity (Liang et al., 2024; Lin, 2025; Pang et al., 2025; Salvagno et al., 2023; Wallwork, 2024; H. Wu et al., 2023). Applications range from summarising related work to translating technical language into more accessible prose, thereby accelerating the writing process. Some platforms extend these capabilities toward fully automated drafting and documentation, including code annotation (Binz et al., 2025; Dou et al., 2024). Scientists report that these tools help in composing readable introductions, summarising related work, or converting technical jargon into clearer language (Yuhan Liu et al., 2025). The overall result is often quicker writing and enhanced clarity, although authors should thoroughly fact-check AI-generated text for accuracy (Augenstein et al., 2024; Dierickx et al., 2024; Giarelis et al., 2024). This trend has been so rapid that it has had an unprecedented impact on scientific writing, surpassing even the effect of major global events on language (Kobak et al., 2025). The academic community is now grappling with policies for disclosure and proper use of AI in manuscript preparation (Bhavsar et al., 2025).

- **Summarisation and translation** – AI is also used to communicate results to broader audiences. For example, some journals and conferences utilise AI summarisers to generate plain-language summaries of technical papers for press releases (Glickman and Zhang, 2024; Markowitz, 2024). Additionally, machine translation powered by AI (e.g. *DeepL*, *Google*'s translation algorithms) enables research written in one language to be more easily understood by non-native speakers (Li et al., 2025; Polakova and Klimova, 2023), breaking down language barriers in science communication and allowing findings to reach all communities.

- **Enhanced visualisation** – Communicating scientific results often involves charts, graphs and images. AI tools can assist in generating more effective visual communications (Dibia, 2023; Maddigan and Susnjak, 2023; A. Wu et al., 2022; Wu et al., 2024), for instance, by automatically selecting the best chart types or even generating schematic diagrams from data. There are experimental systems where a scientist can input raw data, and an AI tool suggests insightful ways to plot it or even creates graphical abstracts (A. Wu et al., 2022).

- **Quality control integrity in publishing** – Alongside generating text and images, AI is utilised by journals and the scientific community to review communications. Plagiarism-detection software has long been used to compare manuscripts against databases (Gupta and Pruthi, 2025; Pudasaini et al., 2024). More recently, AI tools like *Proofig* have been developed to scan submitted papers for image manipulation or duplication, a known issue in some published works (Van Noorden and Perkel, 2023). These algorithms can rapidly compare all figures in a paper to spot if the same microscopy photo has been reused or tampered with. Similarly, there are AI-based tools to try to detect if a piece of text was likely generated by an LLM (Boutadjine et al., 2025; Elkhatat et al., 2023), though reliably detecting AI-written text remains challenging.

2.7.2 CHALLENGES AND LIMITATIONS

A significant concern is the potential for AI models to hallucinate or fabricate citations and textual content, leading to the dissemination of false information (Eger et al., 2025; Elali and Rachid, 2023; L. Huang et al., 2025; Kabir et al., 2025; Monge Roffarello et al., 2025; Walters and Wilder, 2023). AI-generated text can be challenging to distinguish from human-written text and may bypass plagiarism detectors (Gupta and Pruthi, 2025; Weber-Wulff et al., 2023). Over-reliance on AI can lead to superficial or banal writing that lacks creativity and critical analysis (Buriak et al., 2023; Jaakkola, 2024; Zhai et al., 2024). Frequent use of LLMs might lead to similar paragraph structures and writing styles across papers (Bao et al., 2025; Geng and Trotta, 2024; Melliti, 2024; Muñoz-Ortiz et al., 2024; Wenger and Kenett, 2025). Moreover, ethical considerations require transparent documentation of AI use. In peer review, AI can introduce biases, may not be accurate in assessing research quality, and can be exploited for generating irrelevant comments (Mollaki, 2024; Pataranutaporn et al., 2025; Seghier, 2025; Yang et al., 2025).

2.8 Build scientific community

Science is fundamentally a social endeavour, where researchers collaborate, share knowledge, peer review each other's work, and build on collective insights. AI is starting to play a role in shaping and supporting the communities and networks of science, such as by facilitating knowledge sharing and interdisciplinary work within and across scientific fields (Beck et al., 2022; Berens et al., 2023; X. Hu et al., 2025; Lu, 2024; Xie et al., 2024).

2.8.1 RECURRING AND EMERGING TRENDS

- **Collaboration networks and partner matching** – The same network analysis and recommendation algorithms that tech companies use can be applied to academic data (papers, citations, authors) to identify potential collaborations (Lathabai et al., 2022; C. Liu et al., 2024). By analysing publication and citation networks, AI can identify emerging research groups, interdisciplinary connections, or even recommend mentors and mentees (Resce et al., 2022). For instance, if two labs in different countries are working on complementary aspects of a problem, AI might flag this connection, helping to form new collaborations.

- **Knowledge repositories and shared databases** – AI is enabling more dynamic and structured ways for communities to share knowledge. One example is the Open Research Knowledge Graph (ORKG) (Auer et al., 2021), a platform that uses AI/NLP to convert published findings into a knowledge graph of concepts, methods, and results. Over the last five years, ORKG has grown into a *vibrant platform that enhances the accessibility and visibility of scientific research*, by turning unstructured papers into a structured, queryable knowledge base (Oelen et al., 2024; Stocker et al., 2023). This allows communities to contribute and curate knowledge in a collective resource, making it easier to find related work and compare results. Such AI-curated knowledge graphs essentially serve as living review papers that the community builds together, aided by machine extraction of facts and relationships (Auer et al., 2025). This strengthens the community's shared understanding and helps avoid duplication of effort.

- **Citizen science and public engagement** AI is also influencing how wider communities, including non-professional scientists, participate in research. In *citizen science* projects, volunteers team up to collect or classify data (e.g. identifying galaxies, transcribing texts, monitoring wildlife) (Abdul-Rahman et al., 2025; Fortson et al., 2024). AI is being integrated to support these volunteers, for example, by pre-filtering data so that humans focus on the most interesting cases, or by validating and aggregating volunteer contributions (Kumar, 2025). For instance, on the popular *Zooniverse*[19] citizen science platform, AI image classifiers help direct volunteers to images likely containing the objects of interest (such as rare galaxies), making their efforts more impactful. This synergy allows citizen scientists to achieve more, building a community where human intuition and AI efficiency combine. Moreover, AI chatbots and assistants can answer basic questions for community members, lowering the barrier for public participation in scientific discussions (Kajiwara and Kawabata, 2024).

- **Peer review and community evaluation** – The peer-review process, a cornerstone of scientific communities, is beginning to use AI for support, as previously highlighted. Some journals and conferences deploy NLP tools to match manuscripts with suitable reviewers by analysing the content of submissions and the expertise of reviewers. Additionally, AI text analysis can help spot potential issues in submissions (such as plagiarism checks, scanning for statistical errors or unusual reporting patterns that might indicate problems). While the final judgment is left to human reviewers and editors, AI can lighten their load by catching obvious problems and allowing them to focus on deeper content issues (Alnaimat et al., 2025; J. Q. J. Liu et al., 2024).

- **Community involvement and governance** – The academic community needs to prepare for the shift brought by AI and lead the discussion on how AI should be integrated (Chakravorti et al., 2025; Kohnke et al., 2025; McDonald et al., 2025; Vieriu and Petrea, 2025; Walter, 2024). This involves actively engaging with tools, following advancements, contributing to development, and defining appropriate use to ensure AI aligns with rigorous scientific principles (Huerta et al., 2023). Collaboration between social scientists and AI researchers is considered essential. There is a call for collaborative efforts among researchers, policymakers, and generative AI companies to create *living guidelines* for responsible AI use.[20] Policies on AI use should be developed in collaboration with experts and stakeholders like peer reviewers and program chairs. Large systematic review institutions could officially make AI part of practice by providing guidelines.

2.8.2 CHALLENGES AND LIMITATIONS

The rapid pace of AI development presents a challenge for policy and regulation. There are risks when using tools created by private companies,[21] as their goals may diverge from science and their models can be less transparent than academic ones (He et al., 2023; Pal, 2023; Resnik and Hosseini, 2025). The increasing

[19] https://www.zooniverse.org/.

[20] https://european-research-area.ec.europa.eu/news/living-guidelines-responsible-use-generative-ai-research-published.

[21] https://www.nature.com/articles/d41586-024-02757-z.

autonomy of AI systems necessitates clear accountability and responsibility for developers, researchers, and users (Cano, 2025). Distinct approaches and focus areas exist between the AI and science communities regarding AI integration (Zhang et al., 2021). Training the next generation of scientists requires not only technical education but also exposure to scholarship in Science and Technology Studies (STS), social epistemology, and philosophy of science to navigate AI's epistemic risks (Alvarado, 2023; Coeckelbergh, 2025).

3.

DEEP DIVES

The breadth and diversity of the scientific community render the definition of the scientific method itself difficult to pinpoint. Indeed, what is understood as the use of Artificial Intelligence (AI) can be multi-faceted. In this context, to decipher the intricacies of AI as a tool within the context of the scientific method, an examination of different domains is pertinent to be performed. Across the numerous different research areas, specific deep dives were identified and chosen for their distinct ways of using AI as part of experimentation. The three chosen domains were selected from among those that, in recent years, have garnered both significant attention and interest, owing to technological advancements in the field, and significant funding for research progression:

- Protein structure prediction (Section 3.1)
- Material discovery (Section 3.2)
- Ancient site discovery and virtual restoration of inscriptions (Section 3.3)

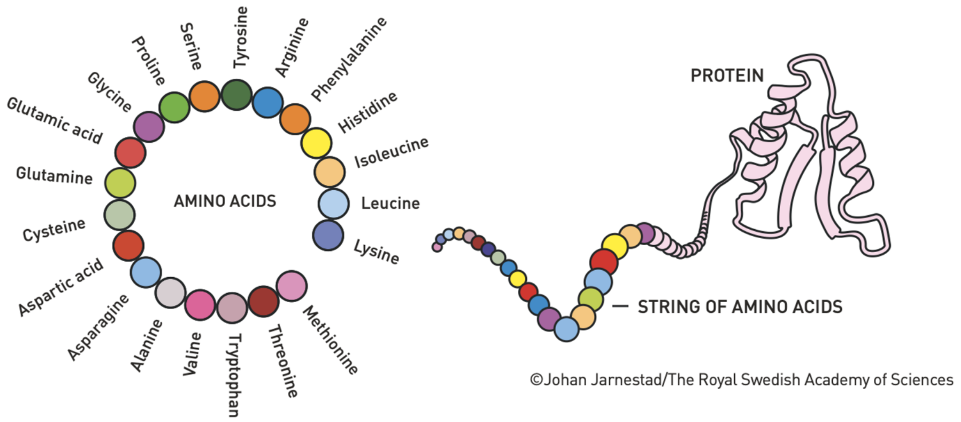
The inclusion of three distinct research fields not only underline how the same technology can be modified and applied across different application areas and disciplines but also showcase common needs of the scientific community. Within each of these domains, the field is described along with its *challenges* and *opportunities*, with presentations of *model architectures*, *databases*, and *infrastructural needs*. The analysis is complemented by an investigation, within each research area, of the *AI usage*, intended as the engagement of the scientific community with the technology, through publications of academic contributions, including journal articles, conference papers and literature reviews. This analysis follows the methodology established by the divinAI project (see Section 1.3).

3.1 Protein structure prediction

In 2024, the Nobel Prize in Chemistry recognised a scientific milestone in the structural biology field with transformative potential for life sciences: the accurate prediction of the three-dimensional (3D) structure of proteins from their string of amino acids, powered by AI.[22] Structural biology is the study of how biological molecules such as proteins and DNA are built and shaped in three dimensions, much like the architecture of tiny molecular buildings. Determining the 3D structure, or simply the structure, of a protein is crucial, as this 3D conformation enables proteins to perform essential functions within living organisms (Martin et al., 1998). Knowing the structures of proteins allow us to understand how they work, or fail to work, in health and disease (Dobson, 2003), enabling the development of targeted therapies and more effective biomedical interventions (Wei and McCammon, 2024).

22 https://www.nobelprize.org/prizes/chemistry/2024/press-release/.

Figure 17. A protein can consist of everything from tens of amino acids to several thousand. The string of amino acids folds into a three-dimensional structure that is decisive for the protein's function.



Source: Johan Jarnestad, Popular information. NobelPrize.org. Nobel Prize Outreach 2025.

For years, , determining the structure of a single protein could take months or even years of laboratory work using complex and expensive techniques like X-ray crystallography or nuclear magnetic resonance spectroscopy (Ledford, 2010). Today, thanks to AI, these structures can be predicted with remarkable accuracy and speed. This revolution was catalysed by the convergence of a long-standing, open, and collaborative initiative known as the Critical Assessment of Structure Prediction (CASP), a global benchmarking effort that, since 1994, has brought together researchers to test and improve protein structure prediction methods (Moult et al., 1995), with rapid advances in computing technology and the accumulation of comprehensive experimental protein data.

In 2020, a major leap occurred when AlphaFold2, developed by DeepMind (a Google subsidiary), dramatically outperformed previous approaches, achieving around 90% accuracy (near-experimental accuracy) in predicting a challenging set of protein structures (Kryshtafovych et al., 2021). By the next round in 2022, most successful participating teams had adopted AlphaFold2-based methods, marking a turning point in the field (Kryshtafovych et al., 2023). In parallel, these advances also enabled more accurate design of synthetic proteins (i.e. engineered molecules with new or enhanced functions), opening new possibilities in therapeutic innovation and biotechnology (Anishchenko et al., 2021).

In a remarkably short span of time after AlphaFold2's debut, the Nobel Prize recognised both the developers of AlphaFold2 (Demis Hassabis and John Jumper) alongside David Baker, a pioneer in protein engineering, for their complementary contributions to structural prediction and synthetic biology (Callaway, 2024). Together, their work is driving a new era in biomedicine, accelerating discovery, enhancing our understanding of disease mechanisms, and unlocking innovative biotechnological applications (Kovalevskiy et al., 2024; Varadi and Velankar, 2023). While the benefits are clear, this shift also presents challenges. These include growing infrastructure demands resulting from the increasing complexity and size of AI model architectures, and the need for extensive, high-quality data.

This deep dive offers an overview of how AI is accelerating scientific progress in structural biology, highlighting both the opportunities and limitations as well as challenges that come with this rapid transformation.

3.1.1 DATA, MODELS AND INFRASTRUCTURE

The success of the advances in deciphering protein structure prediction is rooted in open science, from the availability of the training data to the open-source offering of models and prediction web servers.

Firstly, the key models that have driven transformational progress in the field are examined. Subsequently, the computational resources that have made large-scale predictions feasible are discussed. Finally, an examination of the datasets that have enabled effective training and validation of these models is conducted.

One of the most notable aspects of recent advances in protein structure prediction is the *diversity of AI architectures* now being applied to this complex scientific problem. In such a scenario, especially since the release of AlphaFold2, an improved version of the original AlphaFold, new strategies have explored a range of Deep Learning (DL) models, each with different strengths depending on the use case, available data, and computational resources. Understanding these architectural differences helps explain why no single model can be considered the 'best' for all situations. For clarity, protein structure prediction models have been grouped into three broad categories:

1. **Evolutionary-based models**.
   Evolutionary-based models, including AlphaFold2 (Jumper et al., 2021),

RoseTTAFold (Baek et al., 2021), and OpenFold (Ahdritz et al., 2024), predict protein structures by combining the target protein's sequence with evolutionary information extracted from Multiple Sequence Alignments (MSAs) and structural templates derived from previously solved 3D structures of related proteins. MSAs capture evolutionary conservation that often reflects key structural features, while structural templates provide concrete 3D examples that constrain predictions. These models leverage transformer architectures (Vaswani et al., 2023), which *excel at modelling sequential data and capturing long-range dependencies*, enabling them to learn complex relationships both within the target sequence and across related sequences in the MSA and structural templates (Jumper et al., 2021). This approach delivers state-of-the-art accuracy not only for proteins with many known related sequences but also for those with limited evolutionary data, as DL can detect subtle motifs and generalise from large training datasets. However, these models are computationally intensive to run; generating MSAs and finding suitable templates require significant compute time and memory.[23]

2. **Protein language models**. A growing class of models known as protein language models, including ESMFold (Lin et al., 2023), OmegaFold (R. Wu et al., 2022), IgFold (Ruffolo et al., 2023), pLM-BLAST (Kaminski et al., 2023), and EMBER3D (Weissenow et al., 2022), have the characteristic of eliminating the need for MSAs. Instead, they treat protein sequences like natural language, using transformer-based architectures trained on large corpora of protein sequences to learn structural patterns directly from sequence data. These models are *significantly faster and easier to scale across large datasets* (Lin et al., 2023), making them ideal for rapid screening or exploratory research. The trade-off is that they typically achieve lower structural accuracy compared to MSA-based models (Lin et al., 2023), although they show advantages for achieving higher accuracy for newly design or orphan proteins, which are poorly represented in the training data of evolutionary-based methods (Chowdhury et al., 2022).

3. **Biomolecular interaction models**. A set of recent models, such as RoseTTAFoldNA (Baek et al., 2021), AlphaFold3 (Abramson et al., 2024), Boltz-1 (Wohlwend et al., 2024) and Boltz-2 (Passaro et al., 2025) expands the prediction task of evolutionary-based models to include molecular interactions between proteins and with DNA, RNA, or small molecules. These models aim to reflect the *more realistic, multi-molecular context in which proteins function*. By predicting the structure of molecular assemblies, they offer valuable insights into biological mechanisms. However, this added complexity comes with higher data and computational demands, and some models in this category are not fully open source (Passaro et al., 2025).

In order to highlight the diversity of approaches per model category, in addition to the models mentioned above, Table 2 showcases a non-exhaustive selection of models from each category, chosen based on their scientific impact, methodological novelty or widespread usage in the research community. While it is beyond the scope of this use case to classify these models as foundation models or not, some insights are provided that might help navigate this aspect. *Evolutionary-based models* are in essence task-specific systems for structure prediction, and the field does not adopt the 'foundation model' terminology, although some authors regard them as 'foundation models' for protein-structure-centric tasks since they can be fine-tuned for downstream applications such as protein design (Jue Wang et al., 2024). *Protein language models*

[23] https://elearning.vib.be/courses/alphafold/lessons/alphafold-on-the-hpc/topic/computational-limits/.

more closely resemble foundation models since they are trained on massive protein sequence datasets in a self-supervised manner and learn general representations applicable to multiple downstream tasks including structure prediction. For *biomolecular interaction models*, while most are task-specific systems, emerging models are becoming more powerful and general, as exemplified by Boltz-2, which distinguishes itself as a general platform for biomolecular modelling, positioning itself as a 'biomolecular foundation model'.

Table 2. Overview of representative AI models for protein structure prediction and related biomolecular modelling. Models are grouped by methodological approach: evolutionary-based models, protein language models, and models targeting biomolecular interactions. Key characteristics, open-source availability, and country/region of the corresponding author (used as proxy for region) are also listed.

Types	Characteristics	Models	Open source	Region	Reference
Evolutionary-based	High accuracy, slower, needs highly curated data	AlphaFold2	Y	UK	(Jumper et al., 2021)
		RoseTTAFold	Y	US	(Baek et al., 2021)
		OpenFold	Y	US	(Ahdritz et al., 2024)
		FastFold	Y	SG	(Cheng et al., 2023)
		LightRoseTTA	Y	CN	(X. Wang et al., 2025)
		ScaleFold	Y	CN	(F. Zhu et al., 2024)
Protein language models	Fast and scalable, less accurate	ESMFold	Y	US	(Lin et al., 2023)
		ESM2-15B	Y	US	(Lin et al., 2023)
		ProstT5	Y	EU	(Heinzinger et al., 2024)
		OmegaFold	Y	US	(R. Wu et al., 2022)
		IgFold	Y	US	(Ruffolo et al., 2023)
		pLM-BLAST	Y	EU	(Kaminski et al., 2023)
		EMBER3D	Y	EU	(Weissenow et al., 2022)
		RoseTTAFoldNA	Y	US	(Baek et al., 2024)
Biomolecular interaction	Broader biological scope, more data complexity, computationally intensive	AlphaFold3	N	UK	(Abramson et al., 2024)
		Chai-1	Y	US	(Boitreaud et al., 2024)
		AlphaFold-Multimer	Y	UK	(Evans et al., 2022)
		Boltz-1	Y	US	(Wohlwend et al., 2024)
		Boltz-2	Y	US	(Passaro et al., 2025)

Source: JRC's own elaboration.

After a comprehensive examination of recent advances in protein structure prediction models, attention now turns to the computational infrastructure that facilitates these developments. It is important to distinguish between the infrastructure required to train these models and that needed to run them once developed. Training DL models for protein structure prediction, given the complexity of the problem, which is considered Nondeterministic Polynomial-hard (Rosignoli et al., 2025), demands access to high performance computing environments equipped with powerful CPUs, advanced GPUs, and the capacity to process massive datasets efficiently (Cheng et al., 2023). The resource cost of training a model is reflected in the time taken in addition to the number and type of devices used.

Examples of different models and their respective resource cost are shown in Table 3 and Figure 18. A characteristic comparison can be seen between RoseTTAFold and LightRoseTTA. The former, a model of 130 million parameters required an approximate 5760 GPU hours during training, while LightRoseTTA, at a total of 1.4 million parameters was trained at 168 GPU hours (X. Wang et al., 2025). The variability in the hardware and training time for each model reflects not only a financial cost, but also an environmental impact. Thus, there is a further motivation to be targeting the development of models of high performance and smaller size.

In contrast to training, the infrastructure required to use pre-trained models has become increasingly accessible, with web servers able to run models even for users with limited local computing power. This has resulted in the creation of *model-as-a-service*, such as the case of AlphaFold2 instance provided by the Centre for Education, Research and Innovation in Information and Communication Technologies (CERIT) in the Czech Republic,[24] ColabFold[25] or AlphaFold3 server.[26] From a user's perspective, a laptop can submit jobs via a web UI; however, the compute and memory footprints are still borne by the remote infrastructure. Consequently, accessibility for end-users and the need for substantial investment are not contradictory: these ready-to-use services depend on continued expansion of GPU-rich HPC/AI facilities to meet growing inference demand in science.

[24] https://www.cerit-sc.cz/infrastructure-services/tools-and-applications/alphafold.

[25] https://colab.research.google.com/github/sokrypton/ColabFold/blob/main/AlphaFold2.ipynb.

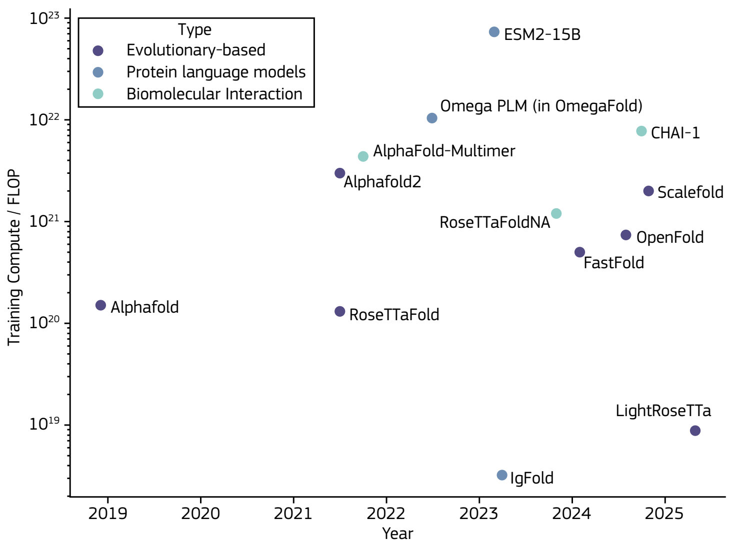
[26] https://alphafoldserver.com.

Table 3. Comparative overview of selected protein structure prediction models. This illustrates the range from large-scale, resource-intensive models (e.g. AlphaFold) to lighter, more accessible implementations (e.g. LightRoseTTA), highlighting differences in hardware configurations, total training time, and compute resources consumed.

Models	Hardware	Training time	Resource
AlphaFold2	128 x TPUv3	11 days	33,792 TPU hours
OpenFold	128 x NVIDIA A100 GPUs	8.39 days	25,774 GPU hours
FastFold	256 x NVIDIA A100 GPUs	2.81 days	20,738 GPU hours
RoseTTAFold	8 x NVIDIA V100 GPUs	30 days	5,760 GPU hours
LightRoseTTA	1 x NVIDIA RTX 3090 GPU	7 days	168 GPU hours

Source: JRC's own elaboration based on literature contributions (Cheng et al., 2023; X. Wang et al., 2025).

Figure 18. Computation costs for various protein structure prediction models, defined in FLOPs. For models with available hardware information during training, FLOPs were calculated using the method described by Epoch AI's 'Data on AI Models'.[27]

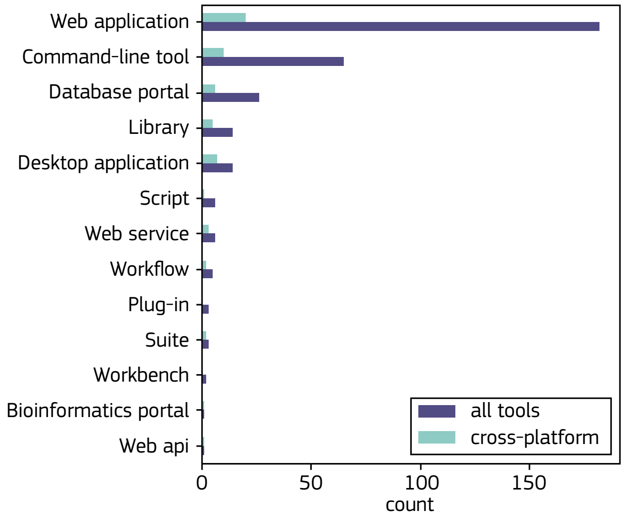


Source: JRC's own elaboration based on Epoch AI's 'Data on AI models'.

[27] https://epoch.ai/data/ai-models.

When software is released to the public, three key criteria are considered: **accessibility**, **user-friendliness**, and **ease of understanding**. An analysis of the choices made in distributing protein structure prediction tools (Figure 19) was conducted, drawing on insights from Rosignoli et al. (2025). It was found that Graphical User Interfaces (GUIs), designed with UX/UI best practices, significantly enhance accessibility and ease of use. The analysis revealed that models were most often released via web servers, which offer high accessibility but may limit user control. Command-line tools remain popular due to their flexibility and efficiency in batch processing. In contrast, desktop applications are less frequently used, likely due to cross-platform compatibility issues, even though they provide greater control and independence from server-based limitations.

Figure 19. Overview of tools retrieved from 'Protein folding' topic tag in the Bio.Tools database.



Source: JRC's own elaboration based on data from the Bio. Tools database (Ison et al., 2019).

Beyond computational infrastructure, the quality and accessibility of data are key to the performance of protein structure prediction models. Two main types of data underpin this field: experimental data and predicted data (see Table 4). Experimental data, derived from real-world measurements, forms the foundation for model training, validation, and benchmarking.

Table 4. Overview of key datasets relevant to protein folding and structure prediction. Where available, the size of each dataset is an estimation.

Type	Dataset	No. of protein structure	Size	Description	Base	Reference
Experi-mental	PDB	240K	60 GB	3D structure	Global	(Ormö et al., 1996)
	Pfam (v.33.1)	47M	141 GB	Protein families	UK, EU	(Paysan-Lafosse et al., 2025)
	SCOP	N/A	N/A	Structural classification	UK, EU	(Andreeva et al., 2020, 2014)
	SCOPe	N/A	86 MB	Extended version of SCOP	US	(Chandonia et al., 2022; Fox et al., 2014)
	CATH43	N/A	1073 MB	Hierarchical classification	UK	(Sillitoe et al., 2021)
Predicted	Alpha-FoldDB	214M	23 TB	3D structures	UK	(Jumper et al., 2021; Varadi et al., 2022a)
	ESM Metagen-omic Atlas	772M	15 TB	Metagenomic protein structures	US	(Lin et al., 2023)

Type	Dataset	No. of protein structure	Size	Description	Base	Reference
Both	UniProtKB/ Swiss-Prot	500K	590 MB	Knowledge base	UK, EU, CH, US	(The UniProt Consortium, 2025)
	UniProtKB/ TrEMBL	2.29M	146 GB	Knowledge base	UK, EU, CH, US	(The UniProt Consortium, 2025)
	UniRef100	314M	76.9 GB	Clustered sets of sequences	UK, EU, CH, US	(The UniProt Consortium, 2025)
	UniRef90	150M	34 GB	90% identity	UK, EU, CH, US	(The UniProt Consortium, 2025)
	UniRef50	53M	10.3 GB	50% identity	UK, EU, CH, US	(The UniProt Consortium, 2025)
	UniParc	5.28M	106 GB	Sequence	UK, EU, CH, US	(The UniProt Consortium, 2025)
	BFD	2.5B	272 GB	Sequence profile	KR	(Jumper et al., 2021; Steinegger et al., 2019; Steinegger and Söding, 2018)
	MGNify	2.4B	165 GB	Microbiome-derived sequence data and predicted structures	UK	(Richardson et al., 2023)
	Uniclust30	N/A	6.6 GB	Clustered protein sequences	EU	(Mirdita et al., 2017)

Source: JRC's own elaboration based on literature contributions (Hu et al., 2022).

Significant efforts by the scientific community have led to the collection and curation of this data into publicly available repositories, most notably the Protein Data Bank (PDB)[28] (Ormö et al., 1996) a database of experimentally determined protein structures, which contained nearly 240,000 structures as of June 2025. The database is maintained by the Worldwide Protein Data Bank (wwPDB), whose members are; the Research Collaboratory for Structural Bioinformatics Protein Data Bank (RCSB PDB), which acts as the maintainer of the archive, the Protein Data Bank in Europe (PDBe), the Protein Data Bank Japan (PDBj), the Biological Magnetic Resonance Data Bank (BMRB) and the Electron Microscopy Data Bank (EMDB). The experimental data from the research community is sent to one of the member organisations and processed at that level. Following processing, the data reaches the archive-maintaining organisation, which makes the data available to all users.[29] This makes the wwPDB a 'data publisher' model for AI-ready science: it aggregates primary experimental results, standardises metadata/formats and provides stable, public access, which are practices that have accelerated structure-aware AI.

Among the key partners of the PDB is the European Molecular Biology Laboratory (EMBL),[30] a leading life sciences research organisation. It focuses on fundamental research, technology development, and training to advance understanding of biological systems. The EMBL plays a fundamental role in data activities, being a key partner of the most important scientific datasets, like the PDB, fundamental for the development of AlphaFold2

[28] https://www.ebi.ac.uk/pdbe/.

[29] https://www.ebi.ac.uk/pdbe/news/wwpdb-charter-full-and-associate-members/.

[30] https://www.embl.org/.

(Jumper et al., 2021). The EMBL embraces AI as a powerful tool that is and will continue to fast-track scientific discovery and it has developed its own AI strategy. Similarly, other research institutions are also leveraging AI to drive innovation in biomedicine. For instance, AITHYRA[31] is a pioneering research institute established by the Austrian Academy of Sciences to advance biomedicine through artificial intelligence. Launched in 2024 with a €150 million grant from the Boehringer Ingelheim Foundation, it aims to revolutionise disease understanding and treatment development by integrating AI early in biomedical research processes. The institute emphasises open-access principles, ensuring its research data is freely available to the global scientific community.

In addition, large biological sequence databases, such as UniProt (The UniProt Consortium, 2025), MGnify (Richardson et al., 2023) and Big Fantastic Database (BFD) (Jumper et al., 2021; Steinegger et al., 2019; Steinegger and Söding, 2018), play a crucial role by providing the amino acid sequences that serve as input for protein structure prediction models. UniProt is coordinated by the EMBL's European Bioinformatics Institute (EMBL-EBI) in the UK, the Swiss Institute of Bioinformatics (SIB) in Switzerland, and the Protein Information Resource (PIR) in the US. MGnify is also based at the EBI in the UK and focuses on metagenomic data analysis. The BFD repository, although not tied to a specific location, is a collaborative effort supporting large-scale computational projects with extensive sequence data. Unlike the PDB, which focuses on experimentally determined protein structures and often involves meticulous manual curation, the sequence databases may not always undergo the same level of manual oversight. Instead, they rely heavily on automated processes to compile vast amounts of sequence data from various sources. This approach allows them to quickly incorporate new sequences, keeping pace with the rapid advances in sequencing technologies and the ever-expanding volume of genomic data. However, it also means that the data might include redundancies or errors that require further validation and curation. Despite these challenges, these sequence repositories are invaluable resources for researchers, offering a comprehensive and up-to-date collection of amino acid sequences essential for advancing computational biology and bioinformatics.

With the advent of AlphaFold2 and other recent methods (see Table 2), it is now possible to generate structural predictions at scale, including for many proteins that previously lacked resolved structures. For example, the AlphaFold Protein Structure Database (Jumper et al., 2021), developed by DeepMind and EMBL-EBI, includes over 200 million predicted protein structures, covering nearly all proteins from 48 species. It is important to know that they vary in quality (Akdel et al., 2022), and are not always a substitute for high-resolution experimental data. However, they significantly complement experimentally determined structures, narrowing the structural knowledge gap, especially for the human proteins (Porta-Pardo et al., 2022). Another example is the ESM Metagenomic Atlas (Lin et al., 2023), which contains over 772 million predicted structures. The ESM Metagenomic Atlas is hosted by Meta, in the US. Given the diverse array of datasets and databases available, standardisation becomes essential. The 3D-Beacons initiative, hosted by EMBL-EBI, unifies predicted and experimental structure models from various of the aforementioned sources (PDBe, AlphaFold DB, SWISS-MODEL, among others) into a common framework with standardised formats and open access under the Creative Commons Attribution 4.0 license (Varadi et al., 2022b). This simplifies integration and broadens usability across research domains.

Managing these large datasets requires robust infrastructure. When hosted on-premises, High-Performance Computing (HPC) clusters are often necessary, but public access to major databases helps reduce technical barriers. Many of these databases are open to community contributions, encouraging collaboration and continuous improvement.

[31] https://www.oeaw.ac.at/aithyra.

The illustrated recent advances in protein structure prediction illustrate how open science, cutting-edge AI models, and shared infrastructure can drive rapid progress in a complex scientific domain. These developments highlight the strategic value of investing in open, interoperable AI ecosystems for scientific innovation. It is crucial that these developments continue to advance openly, ensuring the pace of scientific progress.

3.1.2 SAFE AND TRUSTWORTHY AI, ETHICAL CONSIDERATIONS AND CHALLENGES

The use of AI in protein structure prediction demands a high degree of trustworthiness due to its far-reaching implications for drug discovery, therapeutic development, and biological research. As models like AlphaFold2 increasingly influence experimental design (Edich et al., 2022), resource allocation,[32] and even clinical implementation,[33] it is essential to ensure their reliability, reproducibility, and transparency to uphold scientific integrity and public trust.

Unlike many widely used AI systems, such as large language models like ChatGPT, that often lack explicit confidence scores, protein structure prediction tools like AlphaFold incorporate built-in metrics. The most used one is the predicted Local Distance Difference Test (pLDDT), a confidence score that allows users to judge the quality of the results for each amino acid in the protein structure (Jumper et al., 2021). Similarly, protein language models provide log-likelihood or perplexity metrics to assess the confidence of predictions, which have been shown to correlate with the quality of structure prediction (Lin et al., 2023).

Still, even with these metrics, uncertainty estimates remain imperfect, especially for challenging cases that are underrepresented in training data, such as flexible protein regions that lack fixed structure (Alderson et al., 2023) or proteins embedded in cell membranes (Dobson et al., 2023), many of which are clinically relevant. For example, Agarwal and McShan (2024) stated that *current implementations of AlphaFold2 can provide highly accurate working models for most rigid, well-folded globular proteins, but may have issues predicting other classes of proteins*. This limitation usually arises from the difficulty of experimentally obtaining data of protein structures that are not globular, that is, proteins that are not typical, compact, and stable, given the special experimental conditions required to isolate them. Moreover, for available structures, experimental conditions may introduce biases since the high-salt conditions used to stabilise proteins for X-ray and Nuclear Magnetic Resonance (NMR) measurements differ from cellular environments. Proteins can also adopt different conformations depending on their cellular context and functional states, adding variability not captured in static experimental structures. These combined factors, limited experimental data for non-globular proteins and potential biases in existing structures, result in less reliable predictions and confidence estimates for challenging protein classes.

Another source of bias stems from the type of experimental data by organism: the most represented organism is human (31% of the database entries), followed by synthetic constructs and mice (4% each), and the bacterium *E. coli* (3%).[34] This highlights a clear bias in organism representation, which dominates the training data and results in better predictions for these organisms compared to less studied ones. This organism bias becomes even more problematic in the context of data obsolescence. Protein sequences in reference databases like UniProt are regularly updated, whereas models in the AlphaFold DB remain static. Well-studied organisms such as humans and mice, which dominate the training data, show relatively stable protein sequences, with only 2.8% and 2.4% discrepancies, respectively, between AlphaFold

[32] https://techfundingnews.com/50m-for-ai-programmable-biology-latent-labs-led-by-deepminds-alphafold-alumnus-to-design-novel-proteins/.

[33] https://deepmind.google/discover/blog/stopping-malaria-in-its-tracks/.

[34] https://www.rcsb.org/stats/explore/scientific_name_of_source_organism/.

models and current UniProt entries (Tsitsa et al., 2025). In contrast, less intensively studied organisms face significant obsolescence issues. For instance, the zebrafish (*Danio rerio*) shows a 43.4% discrepancy, with nearly half of its AlphaFold models no longer corresponding to current UniProt sequences due to major curation efforts (Tsitsa et al., 2025). These issues can limit researchers with restricted computational resources who cannot continuously update protein structure predictions, forcing them to rely on potentially outdated models that may no longer accurately represent current sequence data.

Despite these biases and limitations, the scientific community has recognised that for many cases, results are highly accurate, enabling reliable predictions and representing a transformative advancement in structural biology (Akdel et al., 2022). Tasks that were previously infeasible, such as predicting proteins with scarce evolutionary information, making them especially difficult, are now achievable in specific contexts (Porta-Pardo et al., 2022), providing researchers with structural insights for thousands of proteins that would otherwise remain experimentally uncharacterised. This represents a paradigm shift: even with imperfect confidence estimates, having reasonable structural predictions for previously inaccessible proteins, including neglected disease targets (Porta-Pardo et al., 2022) and proteins from plant organisms critical for environmental research (Lin et al., 2025), constitutes an enormous step forward for biological research. However, substantial challenges remain ahead, as research has demonstrated that computational models designed to predict molecular interactions require significant improvement before they can effectively identify drug mechanisms of action (Wong et al., 2022), indicating that the path from structural prediction to clinical application involves considerable additional hurdles.

The field is actively working to address these challenges and emerging ones through community-driven initiatives. These efforts, exemplified by the CASP competition (Kryshtafovych et al., 2023), have inspired other scientific communities to adopt similar collaborative approaches, such as the Virtual Cell Challenge (Roohani et al., 2025). Therefore, it is vital that these community initiatives are maintained given the significant positive impact in the structural biology field. Dedicated structural funding is essential to avoid putting scientific progress at risk, as evidenced by the concerning potential end of the long-running CASP competition due to discontinuation of funding.[35]

Beyond improving prediction accuracy, the field is witnessing remarkable advances in protein design capabilities. Generative models are increasingly capable not only of predicting natural proteins but also of designing novel proteins with desired biological functions (Anishchenko et al., 2021). This development holds great promise for accelerating biomedical breakthroughs. However, it also raises serious biosecurity concerns. These same tools could be misused to engineer harmful proteins, evade biosafety filters or support dual-use research.

At present, the risk of misuse is currently limited by the need for users with advanced expertise in both molecular biology and machine learning, as well as by technical limitations (Peppin et al., 2024). For example, current protein structure prediction models do not fully account for real-life factors like chemical changes to the protein, how it folds inside the cell or how it interacts with other proteins. However, these constraints could rapidly diminish as the models become more advanced and more accessible, which is witnessed through the increasingly more central role that LLMs are assuming in instances such as the provision of guidance on protocols. To help mitigate these risks, the developers of AlphaFold3 have implemented technical safeguards and responsible use principles,[36,37] for example, restricting the modelling of certain viral proteins on server-based platforms. Still, locally run or stand-alone versions remain less regulated.

[35] https://www.science.org/content/article/exclusive-famed-protein-structure-competition-nears-end-nih-grant-money-runs-out/.

[36] https://alphafoldserver.com/terms/.

[37] https://alphafoldserver.com/output-terms/.

To proactively address emerging risks, policy frameworks around biosecurity and AI are beginning to take shape. The first International AI Safety Report (Bengio et al., 2025) and the Federation of American Scientists[38] has emphasised the need for clear risk mitigation strategies and biosecurity guidelines in the context of AI. Leading scientists and organisations have also contributed frameworks and raised concerns focused about AI-related biological risks (Bloomfield et al., 2024), including efforts specific to AI-driven protein design (Baker and Church, 2024; M. Wang et al., 2025). Leading tech/ AI companies are also raising concerns and risk levels.[39,40] These initiatives collectively underscore the importance of identifying and mitigating potential risks posed by biomolecular AI models capable of generating novel proteins or other molecules with harmful biological effects.

Managing these risks is particularly challenging because open, global AI and biological research are essential to preserving scientific progress and international collaboration. Among ongoing governance efforts, the EU has taken a leading role.

In summary, given the current state of AI development in the protein field, it is still too early to anticipate all potential risks. However, it is encouraging to see that expert groups and regulators are proactively prioritising this issue. Moving forward, the governance of AI in protein science, and in the life sciences more broadly, must strike a careful balance: enabling innovation while implementing safeguards that are technically sound, internationally coordinated, and responsive to rapidly evolving capabilities and threats.

[38] https://fas.org/publication/bio-x-ai-policy-recommendations/.

[39] https://cdn.openai.com/pdf/18a02b5d-6b67-4cec-ab64-68cdfbddebcd/preparedness-framework-v2.pdf.

[40] https://www.anthropic.com/news/anthropics-responsible-scaling-policy/.

3.1.3 SKILLS AND INNOVATION

The breakthrough of AI-driven protein structure prediction, exemplified by tools like AlphaFold2 and others (see Table 2), illustrates not only the power of algorithms but also the evolving profile of the modern researcher, the AI-powered (protein) researcher. This transformation has important implications for both the skillset needed in research and the broader innovation landscape.

Modern research workflows demand an interdisciplinary blend of expertise. The future-ready AI-powered researcher and teams combines deep domain knowledge in structural biology with proficiency in ML methods, data science, and software engineering. This hybrid expertise enables them to curate complex biological datasets, critically assess model predictions, and adapt AI tools to answer new scientific questions. A striking example of this interplay emerged right after the release of AlphaFold2 results: research groups rapidly integrated advanced DL architectures with structural biology insights, resulting in alternative implementations and open-source projects such as RoseTTAFold and later OpenFold (Table 2). This agility reflects a new generation of researchers capable of understanding both the complexity of AI models and their biological implications.

Yet technical skills alone are insufficient. Successful AI-powered researchers also need a collaborative mindset to bridge computational and experimental disciplines. The impact of this collaborative spirit is exemplified by the partnership between the EMBL-EBI and DeepMind, which led to the co-creation of the AlphaFold Database. EMBL, with over 110 research groups covering the breadth of molecular biology (Stephen Cusack et al., 2021), contributed expertise in data curation and integration, making AlphaFold's predictions broadly accessible and usable within the biomolecular research community. This collaboration not only produced a resource covering millions of protein structures but also advanced data standards and delivered targeted predictions, for example, for 17 organisms on the World Health Organization (WHO)'s list of neglected tropical diseases and for 10 organisms linked to antimicrobial resistance (Stephen Cusack et al., 2021).

In addition, the AI-powered researcher must actively address ethical considerations, data biases, and reproducibility challenges inherent to AI-driven science. While AI research has long explored interpretability and explainability in health (Bertolini et al., 2025), these aspects remain less developed in protein biological contexts, though initial efforts have begun in protein structure prediction (Cheng et al., 2023; Parsan et al., 2025; Vecchietti et al., 2024) and protein design (Hunklinger and Ferruz, 2025; Medina-Ortiz et al., 2025).

In EU, the AI-powered research profile is increasingly supported by initiatives like the European AI Factories,[41] which will bring together data resources, compute infrastructure, and interdisciplinary expertise to help scientists harness AI effectively. Such environments enable researchers not only to apply existing models but also to innovate at the intersection of biology and artificial intelligence, bridging algorithmic development and biological discovery to advance the field even further. Complementary programs, like the MSCA (Marie Skłodowska-Curie Actions) Doctoral Networks and MSCA co-fund actions, are already supporting talent pipelines and collaborations that help researchers apply existing models and innovate at the interface of biology and AI.

The increasing accessibility of AI-powered protein structure prediction models is fundamentally changing the research landscape, fostering new opportunities for collaboration between academia and industry. This shift highlights how a combination of technical breakthroughs, open science principles, and user-friendly tools is accelerating the translation of foundational discoveries into practical applications and amplifying their real-world impact.

The dynamic landscape of startups plays a crucial role in driving the field forward. Startups advance innovation across four key areas: improving precision and prediction through novel algorithms; building collaborative networks with research groups; exploring new therapeutic frontiers; and creating technological synergies by integrating AI with other cutting-edge approaches. Their agility enables them to pivot quickly and establish impactful collaborations. However, they also face challenges, such as the need for significant computational resources and navigating complex regulatory landscapes, areas that can become opportunities for differentiation and optimisation. Examples of such startups include AlphaFold AI; ProteoDesign, which engineers proteins for specific functions; FoldLogic, which addresses protein misfolding; and EnzymeCorp, which focuses on industrial enzyme applications.[42] Horizon Europe contributes to innovation by supporting start-ups through the European Innovation Council (EIC) and AI investments to unlock the power of data and AI for breakthrough innovation, helping translate emerging discoveries into practical applications and products while facilitating the journey from laboratory research to market deployment.

Ultimately, by investing in researchers who are as comfortable working with algorithms as they are interpreting biological meaning, and by fostering environments where AI tools are designed for usability and transparency, the field can ensure that breakthroughs move swiftly and responsibly from computational models to tangible benefits in medicine and beyond. This interdisciplinary fluency is indeed crucial, but equally important is fostering collaboration between computational and biological experts. Since comprehensive expertise across both domains is scarce, creating structured opportunities for sustained partnerships between AI specialists and domain experts becomes essential, which ensures that computational power is paired with deep biological insight.

41 https://digital-strategy.ec.europa.eu/en/policies/ai-factories/.

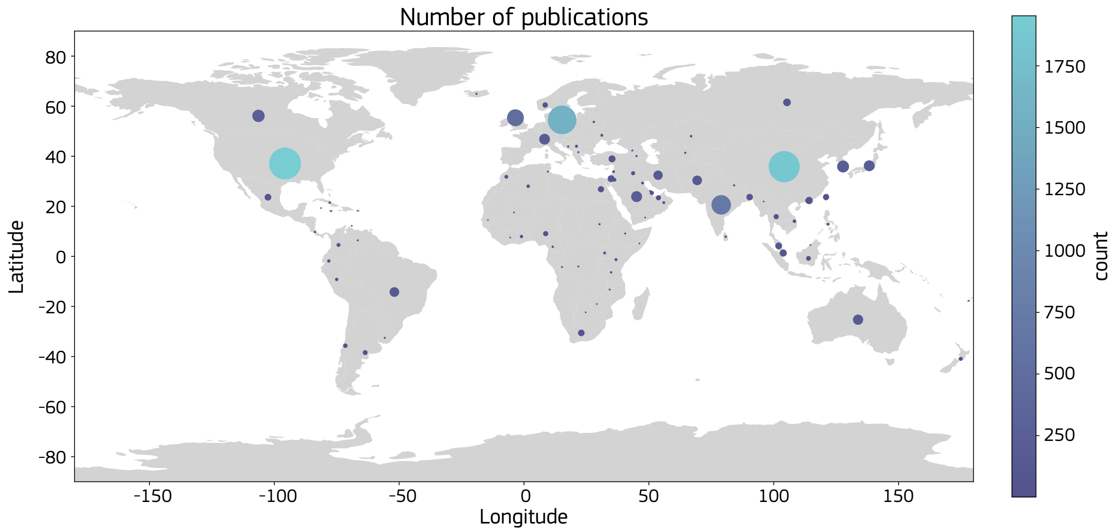
42 https://fastercapital.com/content/Protein-folding-and-aggregation--Startups-and-Protein-Folding--Navigating-Complexity-for-Success.html.

3.1.4 RESEARCH COMMUNITY

Protein structure prediction using AI is recognised as a highly interdisciplinary endeavour, situated at the intersection of computational science, ML, biology, and biophysics. This field has rapidly advanced due to open collaboration among researchers, industry practitioners and government-funded institutions across the globe. To provide a more detailed illustration of the evolution of this research community, an analysis was conducted on 12,307 publications indexed in Scopus between January 2020 and May 2025 and retrieved following the methodology described in Section 1.3. This investigation focused on articles, reviews and conference papers considering the keywords described in Table 1. This quantitative overview highlights the geographical distribution of research activity (Figure 20), the diversity of contributing institutions (Figure 21 and Figure 22), and the collaborative framework that supports scientific progress in this AI-driven domain (Figure 23).

In terms of geographical distribution (Figure 20), the largest number of publications within the defined scope originated from China (2,616; 21.3% of total), the US (2,418; 19.6% of total), EU (2,181; 17.7% of total), India (877; 7.1% of total) and the UK (618; 5% of total).

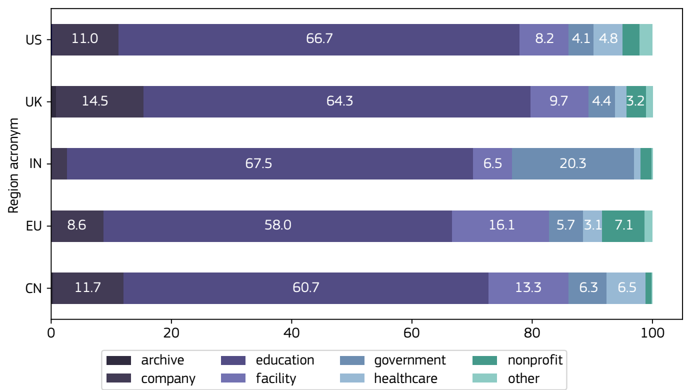
Figure 20. Global geographical distributions of scientific literature contributions related to 'protein structure predictions' topics, indexed by Scopus between Jan. 2020 and May 2025.



Source: JRC's own elaboration based on divinAI methodology.

Publications were produced by a diverse set of establishments, with universities being the predominant contributors, followed by companies and government research facilities (Figure 21). This breakdown underscores the central role of academia in advancing the field, while also highlighting significant engagement from industry and public research institutions. A higher percentage of publications from EU research facilities is observed in comparison to other countries/regions, which highlights the government investment in the studies and the presence at the heart of how research is driven in EU.

Figure 21. Distribution of 'protein structure prediction' literature contributions by type of establishments and geographical area. Considered contributions are indexed by Scopus between Jan. 2020 and May 2025.

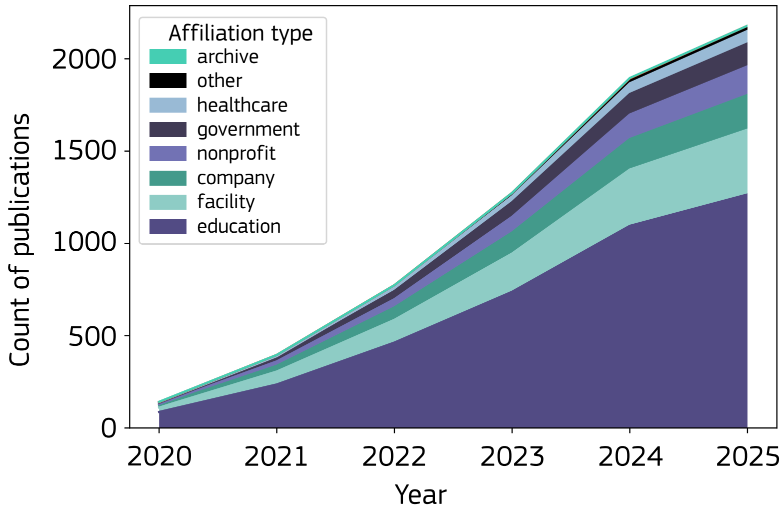


Source: JRC's own elaboration based on divinAI methodology.

Taking a closer look at the synergies within Europe, a more granular analysis of the data shows a spread of research contributions across different types of establishments (Figure 22). It is evident that publications are consistently driven by educational establishments. This distribution reflects that open-source datasets and pre-trained models have significantly lowered the barriers to entry, enabling research groups from both larger and smaller institutions to engage in an area that previously demanded substantial computational costs. For example, AlphaFold2 was reported to have been trained on 128 TPUs (Cheng et al., 2023), greater than the combined resource intensity of all the competing teams in the CASP's 14th edition in 2020.[43] Thanks to community resources and shared infrastructure, many European laboratories are now able to participate actively, despite having smaller local compute budgets. In terms of private sector, which includes start-ups and scale-ups, an increase of 325% can be observed in the period 2020-2024.

[43] https://www.blopig.com/blog/2020/12/casp14-what-google-deepminds-alphafold-2-really-achieved-and-what-it-means-for-protein-folding-biology-and-bioinformatics/.

Figure 22. Amount of 'protein structure prediction' literature contributions across EU countries by type of establishments over the years from Jan. 2020 to May 2025. Considered contributions are indexed by Scopus.



Source: JRC's own elaboration based on divinAI methodology.

Beyond the diversity of institutions, *collaboration* also plays a critical role. Patterns were examined by analysing co-authorship networks based on a selected set of publications. The analysis indicates that the largest cross-regional partnership is identified between the EU and the US, as well as between the US and both the EU and China (Figure 23(a)). A finer level of granularity was then applied to investigate collaboration patterns within the EU by type of establishment (Figure 23(b)). Data reveal that the most common connection (among different type of establishments) occurs between educational institutions and either research facilities or companies. Additional, though less frequent, relevant collaborations involve universities working with government bodies, healthcare institutions and non-profit organisations.

Figure 23. Collaboration patterns in 'protein structure prediction' field. Amount of co-authored literature contributions by (a) geographical area and (b) type of establishments (only EU). Considered contributions are indexed by Scopus between Jan. 2020 and May 2025.

Source: JRC's own elaboration based on divinAI methodology.

Protein structure prediction: Take-home messages

- AI has enabled a paradigm shift in structural biology by accurately predicting protein structures with near-experimental precision, a breakthrough recognised by the Nobel Prize in Chemistry in 2024.
- This success is underpinned by a commitment to open science, leveraging publicly available datasets (e.g. PDB, UniProt) and collaborative benchmarking efforts (e.g. CASP).
- While training these models requires significant computational resources, the dissemination of pre-trained models through web tools and shared infrastructures has democratised their use for the wider scientific community.
- Despite these advances, challenges remain in predicting the structures of non-globular proteins and addressing data biases. The dual-use potential of protein design models also necessitates proactive biosecurity and governance frameworks.

3.2 Material discovery

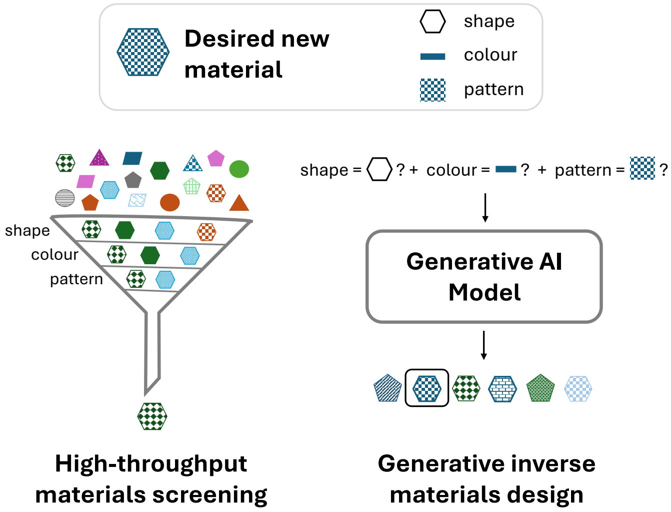
Materials science is undergoing a transformation fuelled by the integration of AI, which is reshaping the landscape of **material discovery**. AI's capability to rapidly accelerate material discovery holds the potential to revolutionise critical areas such as carbon capture (Manikandan et al., 2025), semiconductor design (Raghuwanshi, 2024), energy storage (Zhao et al., 2020), catalysis (Zhao et al., 2019), and biomaterials (Pugliese et al., 2025), paving the way for real-world applications such as more efficient solar panels, improved batteries for electric vehicles, innovative catalysts for emission reduction, and personalised medicine. Therefore, this acceleration is not just a matter of speed. It has profound implications for sustainability, energy efficiency, and the development of novel technologies that could redefine our interaction with the natural world.

Historically, materials discovery has been a labour-intensive and time-consuming process, relying on human intuition, knowledge, and exhaustive trial and error experimentation, with long iteration cycles which limited the number of candidates that could be tested. Thanks to the availability of open material databases (Curtarolo et al., 2012; Kirklin et al., 2015; Scheidgen et al., 2023; Talirz et al., 2020) and recent advances in high-throughput screening (Curtarolo et al., 2013) and AI-based property predictors, it is now possible to screen hundreds of thousands of materials and identify promising candidates in a substantially shorter time.

Despite their advantages, screening-based methods are still fundamentally constrained by the number of known materials. For example, the largest explorations of previously unknown crystalline materials are on the order of 10^7 (Zeni et al., 2025), which is only a small fraction of the number of potentially stable inorganic compounds (Davies et al., 2016). Furthermore, these methods have not yet been shown to be efficiently steered towards discovering materials with specific target properties.

Given these limitations, there has been growing interest in the so-called **inverse material design** approach (property-to-structure), which represent a paradigm shift in material discovery. Rather than starting with a known set of material structures and predicting their properties, the inverse approach begins by defining a set of desired properties (constraints) and seeks to identify the corresponding molecular or extended crystal structures that satisfy them (Park and Choi, 2024) (see Figure 24). Initial approaches to achieving the goal of inverse design were based on global optimisation in the chemical space, for example by using gradient descent (Freeze et al., 2019) or evolutionary methods (Allahyari and Oganov, 2020). However, the most promising approaches are based on generative AI (GenAI) models.

Figure 24. Schematic representation of high-throughput materials screening and generative inverse material design approaches.



Source: JRC's own elaboration, adapted from multiple sources including Park and Choi (2024).

In essence, a GenAI model is trained to map input data, in the form of a numerical tensor that encompasses chemical composition, crystal structure and relevant physical properties of each material, to a latent or feature space represented as a joint probability distribution. This mapping is typically referred to as encoding, and the inverse process is called decoding, which involves transforming from the latent space back into the original data space. Both encoding and decoding models are obtained through learning from a large amount of data. Generation is then the creative process by which novel compounds are produced by sampling from the joint probability distribution and decoding back into the original input space (Park et al., 2024). The most commonly used GenAI model architectures adopted in this domain are:

- *Variational Autoencoders* (VAEs), which combine an encoder that maps input data to a continuous learned latent representation with a decoder that generates samples by drawing from the probability distribution.

- *Generative Adversarial Networks* (GANs), which are based on a generator model that produces synthetic data, and a discriminator model trained to distinguish between real and synthetic data.

- *Autoregressive transformer-based architectures*, including encoder-only and decoder-only attention-based approaches, which are widely used for LLMs.

- *Graph Neural Networks* (GNNs), while not inherently a GenAI model, can be utilised as a component in various generative architectures such as graph VAEs and graph GANs to generalise Neural Networks (NNs) to irregular-shaped graph structures such as chemical graphs of atoms and bonds, 3D structures, or point clouds of atoms, enabling the learning of complex relationships and patterns in material structures.

- *Diffusion models*, emerging architectures that generate sampled materials through a series of iterative stochastic transformations applied to an initial noisy data distribution.

While inverse design represents a central and rapidly advancing area of AI-driven materials discovery, it is worth noticing that it is not the only relevant application. AI is also increasingly used to analyse experimental and simulation data, identify structure–property relationships, and optimise synthesis or processing conditions (Jiang et al., 2025; Li et al., 2020; Pyzer-Knapp et al., 2022; Wang and Zhang, 2025).

3.2.1 DATA, MODELS AND INFRASTRUCTURE

As in other domains, advancements in materials science have been driven by the combination of new computational approaches and open-access datasets. These have been instrumental in material simulation, property prediction, high-throughput screening and, more recently, inverse design. The number of database and dataset resources in materials science is remarkable (Oturak, 2025). Table 5 highlights some of the most relevant resources that are fundamental for material discovery.

One of the most important examples is the Materials Project (Jain et al., 2013) which is a multi-institution, multi-national effort to compute the properties of all inorganic materials and openly provide the data and associated analysis algorithms. The project was established in 2011, and so far (July 2025), they provide more than 200K materials and 577K molecules[44] (see Figure 25). Other well-known databases are AFLOW (Curtarolo et al., 2012), an open-access database of more than 3.5M material compounds with over 743M calculated properties; Novel Materials Discovery (NOMAD) laboratory (Scheidgen et al., 2023), which contains more than 100M high-quality calculations; Open Quantum Materials Database (OQMD) (Kirklin et al., 2015), which provides Density Functional Theory (DFT) calculated thermodynamic and structural properties of more than 1.3M materials; GNoME (Merchant et al., 2023), which released more than 2.2M predicted new crystals, including 380K stable materials; and Alexandria (Schmidt et al., 2023), that contains more than 5M DFT calculations for periodic 3D, 2D and 1D compounds.

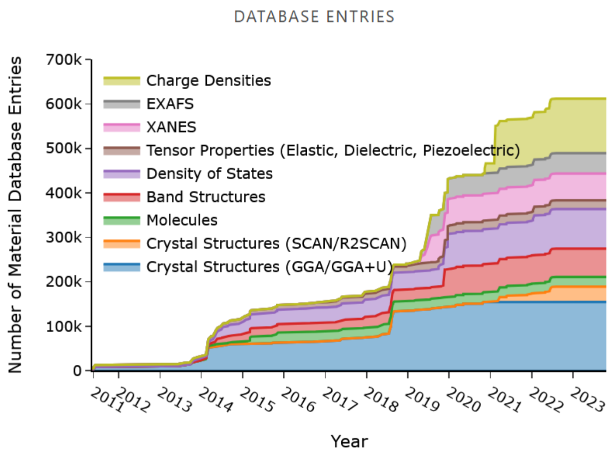
Table 5. Summary of some of the most representative materials science databases and repositories, in chronological order, including database name, type of data (empirical/computational), open-source availability, country/region of the original contribution (used as proxy for region) and literature reference.

Database	Type	Open source	Region	Reference
AFLOW	Computational	Y	US	(Curtarolo et al., 2012)
Materials Project	Computational	Y	US	(Jain et al., 2013)
OQMD	Computational	Y	US	(Kirklin et al., 2015)
NOMAD	Computational	Y	EU	(Scheidgen et al., 2023)
GNoME	Computational	Y	US	(Merchant et al., 2023)
Alexandria	Computational	Y	EU	(Schmidt et al., 2023)

Source: JRC's own elaboration.

[44] https://next-gen.materialsproject.org/.

Figure 25. Number of material database entries provided in the Materials Project since 2011.



Source: The Materials Project (Jain et al., 2013).

Beyond open-access repositories and generative model architectures, several large-scale national and international initiatives have emerged to establish integrated platforms for data-driven materials discovery. In Europe, the DiaDEM project[45] (Digital Discovery platform for Organic Electronic Materials) is developing a platform for the design of organic semiconductors by connecting virtual screening, synthesis planning, and chemical supply chains. The initiative also provides a dedicated dataset[46] with curated molecular structures and optoelectronic properties to support digital screening workflows. Complementing this effort, MaterialDigital,[47] a German initiative, aims to harmonise digital representations of materials by introducing machine-readable standards and common ontologies spanning the entire materials lifecycle, from production to application. The initiative focuses on aligning academic and industrial practices to reduce fragmentation and foster data interoperability.

At the international level, Japan's Materials DX initiative,[48] launched under the national Strategy for Strengthening Materials Innovation Capability, supports the long-term deployment of data-driven approaches through the DxMT programme[49]. This effort seeks to accelerate the development of materials with innovative functions, particularly in strategic domains such as sustainability and carbon neutrality, by integrating accumulated scientific knowledge with AI and data science methods.

These initiatives represent a strategic effort to go beyond datasets, supporting the creation of national-scale platforms that integrate modelling, data infrastructure, and real-world experimentation, thereby reinforcing the role of AI in materials innovation.

GenAI models for inverse material discoveryhave greatly benefited from the availability of these databases. One fundamental aspect that distinguishes between models is the choice of crystal structure representation. At a higher level, representations are categorised into those based on coordinates (e.g. *(x, y, z)* components of the *(a, b, c)* lattice vectors of the unit cell), 3D voxels or point clouds, standardised Crystallographic Information Files (CIF), or other type of representations (e.g. Wyckoff positions). In Table 6, a summary of the most relevant models for inverse material discovery is provided, in chronological order, including the type of structure representation used for the input, the model architecture and whether the model is open source. As can be observed, the most used representation is based on coordinates, with some early examples using voxels and more recent cases using CIF. Regarding model architectures, with few exceptions, the evolution has progressed from early models based on GANs and VAEs to more recent models based on autoregressive transformers and diffusion models. Most of the models are open source, although there are some recent notable exceptions, such as MatterGen (developed by Microsoft) (Zeni et al., 2025). However, while Table 6 presents 'open-source' status as a binary attribute, it is important to acknowledge that the reality is more nuanced. In several cases, including material science research but not limited to it, despite this label, critical

45 https://www.diadem-project.eu/.

46 https://diadem-staging.de/.

47 https://www.materialdigital.de/.

48 https://unit.aist.go.jp/mdx/en/.

49 https://dxmt.mext.go.jp/en/.

components (e.g. pre-trained weights, training pipelines, fine-tuning protocols or proprietary datasets) are withheld (Persaud et al., 2024; Van et al., 2025). This partial openness can lead to erroneous perceptions of reproducibility and accessibility.

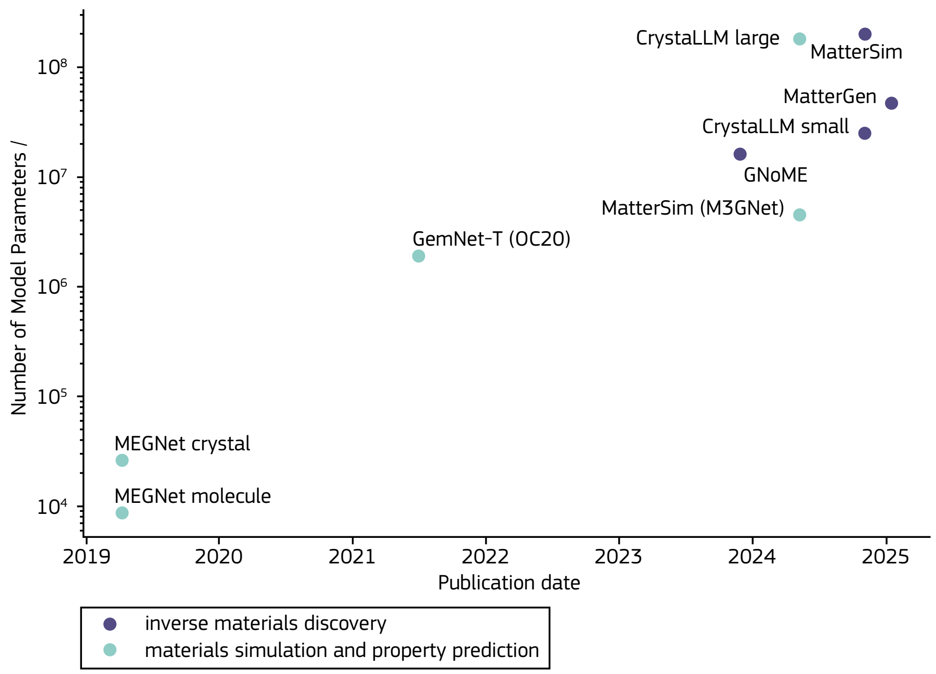
Table 6. Summary of generative AI models for inverse material discovery in chronological order, including model name, input type, model architecture, open-source availability, country/region of the corresponding author (used as proxy for region) and literature reference.

Model	Input type	Model architecture	Open source	Region	Reference
CrystalGAN	Coordinates	GAN	Y	EU	(Nouira et al., 2019)
Crystal-VAE	Voxels	VAE	Y	CA	(Hoffmann et al., 2019)
iMatGen	Voxels	VAE	N	KR	(Noh et al., 2019)
G-SchNet	Coordinates	Generative-SchNet	Y	EU	(Gebauer et al., 2019)
CCCGAN	Coordinates	GAN	N	KR	(Kim et al., 2020)
ICSG3D	Voxels	VAE	Y	UK	(Court et al., 2020)
CCDCGAN	Voxels	GAN	N	EU	(Long et al., 2021)
CubicGAN	Coordinates	GAN	Y	US	(Zhao et al., 2021)
CDVAE	Coordinates	VAE + Diffusion	Y	US	(Xie et al., 2021)
FTCP	Coordinates and structured factors	VAE	Y	SG	(Ren et al., 2022)
xyzTransformer	CIF	Transformer	N	CA	(Flam-Shepherd and Aspuru-Guzik, 2023
UniMat	Coordinate embedded periodic table	Diffusion	N	US	(S. Yang et al., 2023)
GNoME	Coordinates	GNN	Y	US	(Merchant et al., 2023)
Crystal-LLM	Coordinates	Transformer	Y	US	(Gruver et al., 2023)
DiffCSP	Coordinates	Diffusion	Y	CN	(Jiao et al., 2023)
CrysTens	Coordinates	GAN + Diffusion	Y	US	(Alverson et al., 2024)
CristalFormer	Coordinates	Transformer	N	CN	(Cao et al., 2024)
WyCryst	Wyckoff site matrix	VAE	N	SG	(R. Zhu et al., 2024)
CrystalLLM	CIF	Transformer	Y	UK	(Antunes et al., 2024)
MatterGen	Coordinates	Diffusion	N	UK	(Zeni et al., 2025)

Source: JRC's own elaboration based on literature contributions, including (Park et al., 2024).

When it comes to the infrastructure needed to train these models, it is worth noting that the size of these models in terms of the number of parameters is relatively moderate, especially when compared to the size of state-of-the-art GPAI models, such as LLMs. For example, Figure 26 displays the number of parameters for the models with available data. The blue points represent models used for molecular simulation and property prediction. The red points represent models for crystal discovery, including the most advanced ones, such as CrystaLLM and MatterGen, with 200M and 46.8M parameters, respectively. In comparison, recent open-weight 'small' LLMs, such as Mistral Small 3 with 24B parameters, Llama 4 Scout with 17B parameters, or larger models like Llama 4 Maverick with 400B total parameters (based on Epoch AI's 'Data on AI Models'[27]), are between 2 and 4 orders of magnitude larger.

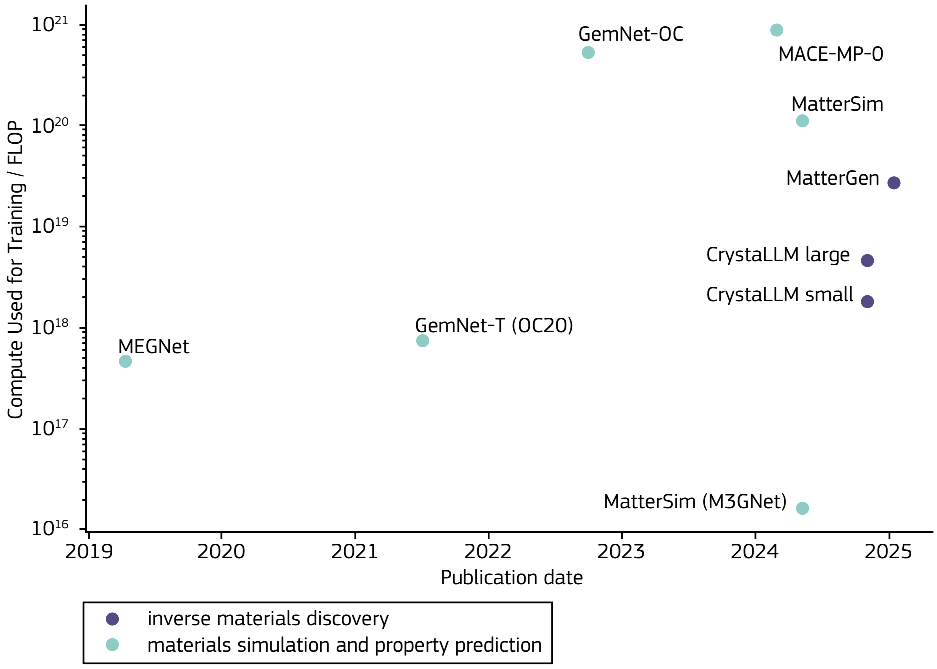
Figure 26. Number of parameters of models for 'materials simulation and property prediction' (blue points) and 'inverse material discovery' (red points).



Source: JRC's own elaboration based on Epoch AI's 'Data on AI models'.[27]

The compute required to train these models is proportional to the size of the model (number of parameters) and the size of the training dataset (number of data points). Some examples of training compute in floating-point operations (FLOPs) are depicted in Figure 27. The most recent models for inverse material design, CrystaLLM (large, transformed-based architecture) and MatterGen (diffusion-based have been trained with a total of 4.6 x 10^{18} and 2.7 x 10^{19} FLOPs, respectively. In comparison, the aforementioned open-weights GPAI models, have been trained with much larger amount of compute, such as 1.15 x 10^{24} FLOPs for Mistral Small 3, or 4.08 x 10^{24} FLOPs for Llama 4 Scout (based on Epoch AI's 'Data on AI Models'[27]). Therefore, in terms of compute, *the difference between the most advance models for materials discovery and the state-of-the-art open-weight GPAI models is approximately 5-6 orders of magnitude*. This is why the current landscape of AI Factories in Europe is, for now, sufficient to meet the training needs of these models in the short to medium term. However, compared to state-of-the-art GPAI models, there is still significant room for improvement, and it is likely that future GenAI models for inverse materials design will be trained with much larger amounts of compute in the future. This increase will likely be driven, at least initially, by the availability of more data, both empirical and computational, as well as the use of larger-scale GenAI model architectures (in terms of the number of parameters), following a similar trajectory to that seen in the development of LLMs.

Figure 27. Training compute (FLOPs) for 'materials simulation and property prediction' (blue points) and 'inverse materials discovery' (red points). Note that training compute are proportional to the number of models parameters and data points used for training.



Source: JRC's own elaboration based on 'Data on AI models'.[27]

Alongside data volume, the quality and consistency of materials datasets are critical for developing reliable AI models. Issues like inconsistent labels, incomplete metadata, and varying simulation parameters can result in misleading predictions and lower generalisability (Himanen et al., 2019). Many existing datasets also lack proper documentation or benchmarking protocols, especially when aggregating from diverse sources (Van et al., 2025). Improving data quality is essential for robust and reproducible materials discovery, and initiatives like the above-mentioned NOMAD are already advancing this agenda. Once a certain level of saturation is reached in terms of data and model scale, algorithmic improvements and data quality will play a crucial role.

3.2.2 SAFE AND TRUSTWORTHY AI, ETHICAL CONSIDERATIONS AND CHALLENGES

The promises of using inverse material design via generative AI to accelerate the discovery of new materials, their properties, and the development of potential new products, patents, and innovation activities are very enticing. While AI models operate in the virtual domain to predict candidate materials with specific target properties, their impact is increasingly amplified when associated with *self-driving laboratories*, i.e. automated experimental platforms that integrate robotics, sensing, and AI to perform synthesis and characterisation with minimal human intervention. These two components are conceptually distinct yet operationally complementary: GenAI proposes hypotheses; self-driving labs test them in closed-loop cycles. Their integration defines a new paradigm of autonomous materials discovery.
In fact, the number of initiatives focusing on intelligent, automated, or self-driving laboratories

has grown significantly in recent years.[50] However, on several occasions, the expectations created by the narratives of some scientific works are perhaps too high and create distortions in both the scientific community and the general public, which may ultimately also impact policy decisions. Both traditional knowledge-led computational materials design and inverse material designs approaches are sometimes criticised for making predictions that are obvious-trivial variants of known systems-, or even erroneous-compounds that are unstable or fail to exhibit the predicted structure/properties (Park et al., 2024).

One example is the case of the A-lab, an autonomous laboratory for AI-driven synthesis of targeted compounds published in Nature (Szymanski et al., 2023), which, just a few days after its publication, and following some criticism regarding the quality of the experimental analysis, had to publish a response[51] with additional data to supplement the original paper, providing evidence that the targeted compounds were indeed successfully synthesised. They declared that *the A-Lab is not intended to replace the materials discovery process with AI agents. Rather, the A-Lab is meant to address a current fundamental limitation in materials science: the number of predicted materials vastly exceeds the capabilities of humans to test experimentally. The focus of our efforts is on the early, exploratory stages of materials discovery: attempting synthesis and gathering data. A human in the loop is still required for the latter stages of discovery.*

Another case that drew criticism was the GNoME tool from Google DeepMind (Merchant et al., 2023), which was announced as a key breakthrough for the discovery of up to 2.2 million new materials. A subsequent study (Cheetham and Seshadri, 2024), however, found scant evidence for compounds that fulfilled the trifecta of novelty, credibility, and utility, concluding that *while the methods adopted in this work appear to hold promise, there is clearly a great need to incorporate domain expertise in materials synthesis and crystallography.*

In this context of high expectations, there have even been extreme cases of possible scientific and academic fraud, such as the recent study on the impact of AI on accelerating materials discovery, patents, and product prototypes (Toner-Rodgers, 2024). The research, presented by a PhD student at MIT, had a significant scientific and social impact, and was covered by some of the most prestigious newspapers and news sections of the most relevant scientific journals (The BS Detector, 2025). Unfortunately, MIT recently stated that, after an internal and confidential review, they have *no confidence in the veracity of the research contained in the paper* (MIT Economics, 2025). The MIT formally requested arXiv and The Quarterly Journal of Economics to withdraw the paper.

These examples highlight the importance of *AI in science research*, particularly in the context of materials discovery, being conducted by *multidisciplinary teams* that incorporate specific domain expertise. Additionally, *full autonomy does not have to be a goal in itself*. GenAI models can still play a valuable role in providing novel suggestions while keeping materials scientists and complementary computational search strategies 'in the loop'. These elements, some of which are key to the EU's approach to Trustworthy AI, are essential to consider in the context of any AI in science strategy that aims to maximise the benefits of AI while minimising unintended and harmful side effects. These potential negative effects can also be minimised by the availability of educational resources, tools, standards, and good practices that can improve expertise and provide the necessary skills to implement AI in various scientific domains.

3.2.3 SKILLS AND INNOVATION

In addition to the general interdisciplinary context that AI in science approaches entails, where domain-specific expertise and skills (in

[50] https://github.com/AccelerationConsortium/awesome-self-driving-labs.

[51] https://www.linkedin.com/pulse/regarding-our-recent-a-lab-article-gerbrand-ceder-0sz6c.

this case, materials science) must be combined with AI-specific profiles typical of the engineering and computer science fields (Sun et al., 2022), it is worth highlighting that in the context of AI in materials science, there are several initiatives and resources (Reeve et al., 2019) that serve as good examples for other research domains, including datasets, simulation tools, tutorials, learning modules, etc. For example, there are a considerable number of initiatives providing open-source tools, tutorials and notebooks on various methods for materials design such as JARVIS-Tools Notebooks (Choudhary et al., 2020), Materials Simulation Toolkit for Machine Learning (MAST-ML) (Jacobs et al., 2020), Machine Learning Lab Module (Afflerbach et al., 2020), or REsource for Materials Informatics (REMI) (NIST, 2025).

Other more structured initiatives include the NOMAD Lab (Scheidgen et al., 2023), which, in addition to serving as a database of resources for training generative models, provides other types of resources such as AI toolkits, an encyclopaedia, tutorials, and documentation. Another relevant initiative is Materials Cloud (Talirz et al., 2020), which is presented as an approach to enable the seamless sharing and dissemination of resources in computational materials science. This platform offers a range of tools, including educational, research, and archiving tools on one hand, and simulation software and services on the other, as well as curated and raw data. All these initiatives are excellent examples of what good practices might look like in the context of improving expertise and providing the skills needed to implement AI in various scientific domains.

In the realm of *innovation*, while some highly significant initiatives in the field of materials science have emerged from major tech companies, such as Meta with OMat24 (Barroso-Luque et al., 2024), Google DeepMind with the GNoME Project (Merchant et al., 2023)[52], or more recently Microsoft with MatterGen (Zeni et al., 2025), the ecosystem of startups specifically focused on leveraging advanced AI tools to enhance materials discovery is nonetheless substantial.[53] The innovative nature of research activities in the materials science domain is strengthened by several aspects that arise from the integration of AI, which, as previously mentioned, not only accelerates the materials discovery process but also makes it much more efficient and sustainable, and focused on addressing very important real-world problems.

3.2.4 RESEARCH COMMUNITY

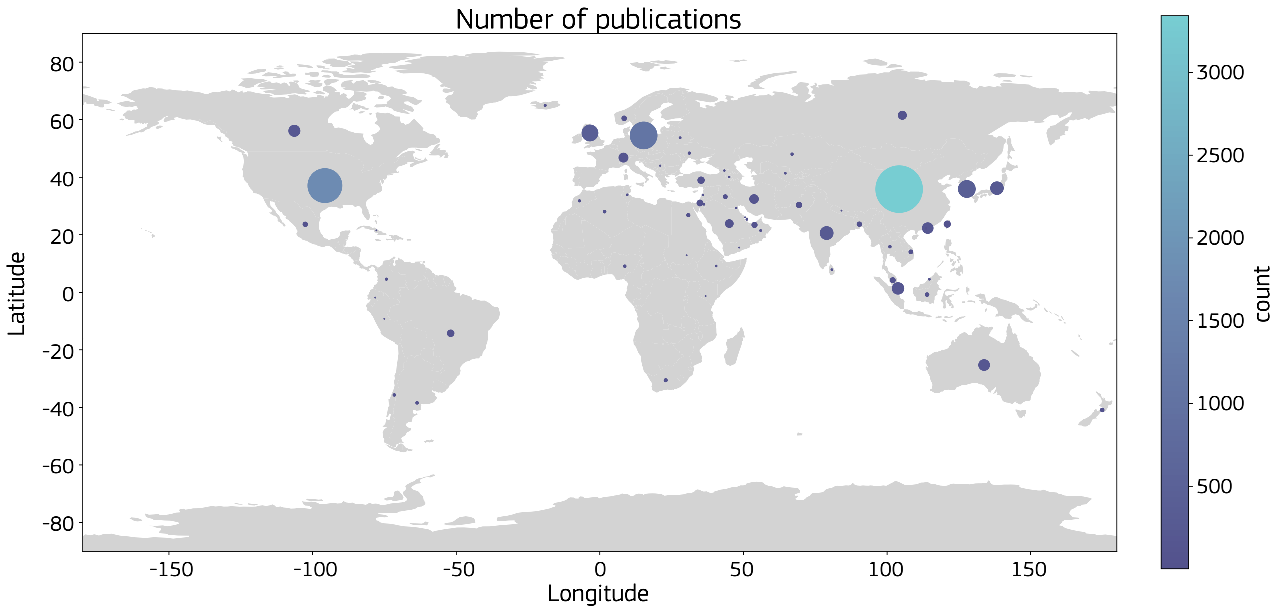
The use of AI for materials discovery has emerged as a rapidly expanding interdisciplinary field at the intersection of materials science, physics, chemistry and ML. This convergence is driven by both technological advancements and the growing recognition of AI's potential to accelerate the discovery of functional materials for energy, electronics, healthcare, and sustainability. The field has benefited from the interaction between domain scientists, data infrastructure projects, and researchers, often supported through publicly funded initiatives and collaborative platforms. To illustrate the evolution and structure of this research community, 7,769 publications indexed in Scopus between January 2020 and May 2025 (retrieved following the methodology described in Section 1.3) are analysed by filtering for articles, reviews and conference papers considering the keywords described in Table 1. This analysis offers a quantitative perspective on the geographical distribution of scientific output (Figure 28), the diversity of contributing institutions (Figure 29 and Figure 30), and the degree of collaboration across countries and sectors (Figure 31).

In terms of geographical distribution (Figure 28), the largest number of publications within the defined scope originated from China (3,342; 35.7%), followed by the US (1,769; 18.9%), the EU (1,098; 11.7%), South Korea (432; 4.6%), the UK (385; 4.1%), India (250; 2.7%), and Japan (241; 2.6%).

[52] Despite the aforementioned recent critique (Cheetham and Seshadri, 2024) of the hyperbolic claims made by Google DeepMind regarding the discovery of 2.2 million structures (*an order-of-magnitude expansion in stable materials known to humanity*), this work remains one of the most impactful in the materials science domain.

[53] '42 Best Startups in Europe to watch in 2025'. https://www.seedtable.com/best-startups-in-europe.

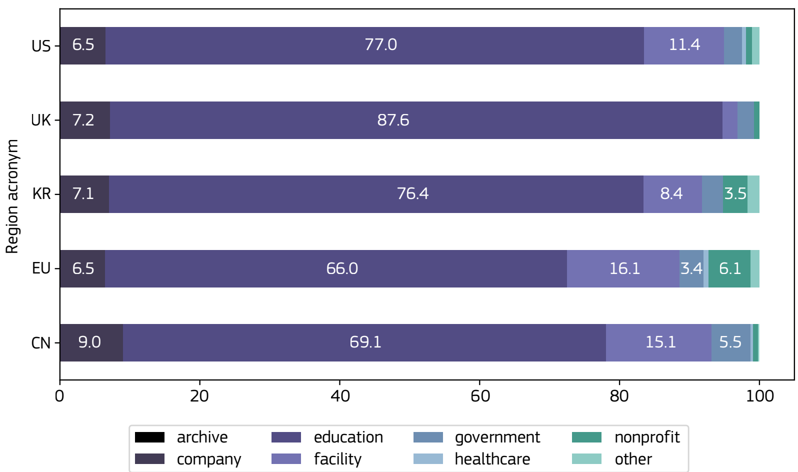
Figure 28. Global geographical distributions of scientific literature contributions related to 'material discovery' topics, indexed by Scopus between Jan. 2020 and May 2025.



Source: JRC's own elaboration based on divinAI methodology.

The publications were produced by a variety of organisations, with universities being the main contributors, followed by research institutions and private companies (Figure 29).

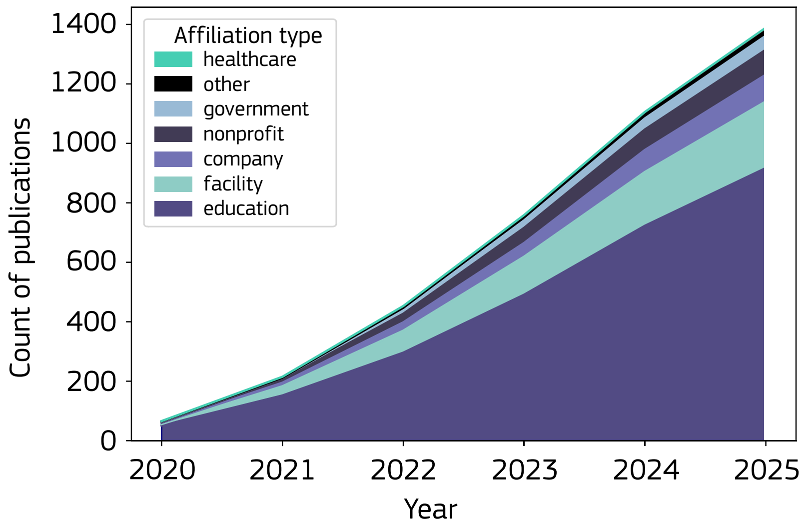
Figure 29. Distribution of 'material discovery' literature contributions by type of establishments and geographical area. Considered contributions are indexed by Scopus between Jan. 2020 and May 2025.



Source: JRC's own elaboration based on divinAI methodology.

In the EU, most research output on AI-driven materials discovery originates from academic institutions, followed by research facilities, private companies, and, to a lesser extent, non-profit organisations, government bodies, and healthcare institutions (Figure 30). This distribution reflects the continued leadership of academia, supported by increasing participation from other sectors. The availability of open datasets, pre-trained models, and shared platforms has made it easier for a wide range of institutions to contribute, including those with more limited computational resources.

Figure 30. Amount of 'material discovery' literature contributions across EU countries by type of establishments over the years from Jan. 2020 to May 2025. Considered contributions are indexed by Scopus.



Source: JRC's own elaboration based on divinAI methodology.

Changing perspective, collaboration networks are also essential to the research landscape. Co-authorship analysis of the selected publications reveal that the most prominent cross-regional collaborations occur between the EU and the US, as well as between the US and China (Figure 31(a)). At a more granular level, collaboration patterns within the EU show that the strongest links are between educational institutions and research facilities, followed by collaborations between academic institutions and companies (Figure 31(b)). Less frequent, but still present, are joint publications involving universities and government bodies, healthcare institutions, or non-profit organisations.

Figure 31. Collaboration patterns in 'material discovery' field. Amount of co-authored literature contributions by (a) geographical area and (b) type of establishments (only EU). Considered contributions are indexed by Scopus between Jan. 2020 and May 2025.

Source: JRC's own elaboration based on divinAI methodology.

Material discovery: Take-home messages

- The field of material discovery is being revolutionised by AI, moving from traditional trial-and-error to novel paradigms, such as inverse material design, which generates novel compounds based on desired properties.
- This progress is powered by a variety of generative AI architectures and a robust ecosystem of open computational material databases, which facilitate high-throughput screening and accelerate innovation in strategic areas like energy and sustainability.
- Despite the significant potential, this research field faces challenges related to inflated expectations, and a critical need for domain expertise to validate AI-generated hypotheses. The computational demands, while moderate compared to LLMs, still require significant investment in research infrastructure.

3.3 Ancient site discovery and virtual restoration of inscriptions

Archaeology is the scientific study of the material remains of past human life and activities. It involves studying ancient artifacts, structures, and landscapes to understand the historical significance of human societies, their formation, and cultural diversity across time. Its mission is to *understand the lives, cultures and behaviours of ancient civilisations and societies that existed long before written records* (Jorge, 2024). Through physical remnants, the development of human civilisation and the evolution of cultures over time can be uncovered. *European archaeology* focuses on the rich and varied heritage of Europe, highlighting both the commonalities and diverse elements that have shaped the continent's history. By uncovering and interpreting materials from different eras, archaeologists contribute to social cohesion by fostering an understanding of our shared, collective past while also celebrating the distinct cultural identities that exist. In recent years, AI has emerged as a powerful tool, offering new methods for analysing archaeological data and enhancing the ability to uncover and interpret historical insights.

The study of inscriptions on long-lasting materials, such as stone, metal or pottery, is crucial for reconstructing historical narratives (Robert, 1961). Through the field of epigraphy, researchers can uncover precious information about ancient languages, cultures, and societies through the deciphering and interpretation of these texts (Bodel, 2012). The precision and depth of historical understanding greatly rely on the efforts of epigraphers, who traditionally use meticulous manual methods to examine and restore inscriptions. European epigraphy holds significant importance due to its extensive record of diverse inscriptions, ranging from ancient Roman and Greek texts to medieval carvings (Millar, 1983). These inscriptions provide a window into cultural, political, and social dynamics, offering critical insights into the rich and varied history of the continent. The study of the artefacts, from objects to structures, is of an inherently interdisciplinary nature, which enables researchers in the field to integrate tools and methods from chemistry, physics, and biology in order to extend their analytical possibilities far beyond those of traditional approaches. When combined with AI, these cross-disciplinary techniques have significantly expanded the capacity of archaeology for discovery and interpretation. Given the breadth of areas in which AI is applied in the diverse field of archaeology, two focus areas were chosen in order to showcase different techniques: **site discovery** and **virtual restoration of inscriptions**. Examples of how AI is being used in these areas are provided below.

By combining advanced techniques in epigraphy with site discovery efforts, researchers can develop a more comprehensive understanding of ancient civilisations, accelerating the discovery process and uncovering new connections between archaeological findings. This integration not only enhances the accuracy and efficiency of archaeological excavations but also contributes to the preservation and interpretation of cultural heritage on a broader scale.

Site discovery is a key area influenced by AI through cutting edge digital sensor data, such as 3D site scans, satellite imagery and environmental datasets. This wealth of data, combined with AI algorithms, enables archaeologists to identify previously unknown sites, classify artefacts and reconstruct patterns of ancient human behaviour (Gattiglia, 2025). In Copán, Honduras, an analysis of airborne Light Detection and Ranging (LiDAR) data using a DL model applied to 3D point clouds was successful in classifying Maya archaeological sites (Richards-Rissetto et al., 2021). Also, in Nazca Desert, Peru, low-flying drones captured high-resolution images of an area spanning 600 km2 of desert. More than 1,300 potential candidates of Nazca culture sites were identified by DL models trained to detect figurative geoglyphs. Subsequent ground verification confirmed 303 novel geoglyphs, almost doubling the number of figures associated with the Nazca culture (Sakai et al., 2024). Lim and Qiu (2023) developed μ-Net, a ConvNeXt-based U-Net architecture that converts muon scattering data into high-resolution three-dimensional density maps, facilitating the non-invasive investigation of ancient structures. Similarly, Benhammou et al. (2025) applied AI-enhanced muon tomography to detect and map subterranean features beneath the City of David heritage site. In forensic archaeometry, Siozos et al. (2021) integrated Laser-Induced Breakdown Spectroscopy (LIBS) with NNs to re-associate commingled human bone fragments in complex burial contexts, achieving high classification accuracy and contributing to more efficient reconstruction of fragmented remains. The above examples show that using AI algorithms to analyse aerial and satellite data can increase the size of archaeological records far beyond what can be achieved through ground surveys alone. Together, these studies exemplify how AI-enabled methods, ranging from ML for peptide analysis, Convolutional Neural Networks (CNNs) for spectroscopic imaging, DL for muon tomography, to NN classification in LIBS, are expanding the interpretative capabilities of archaeology beyond the traditional methodologies.

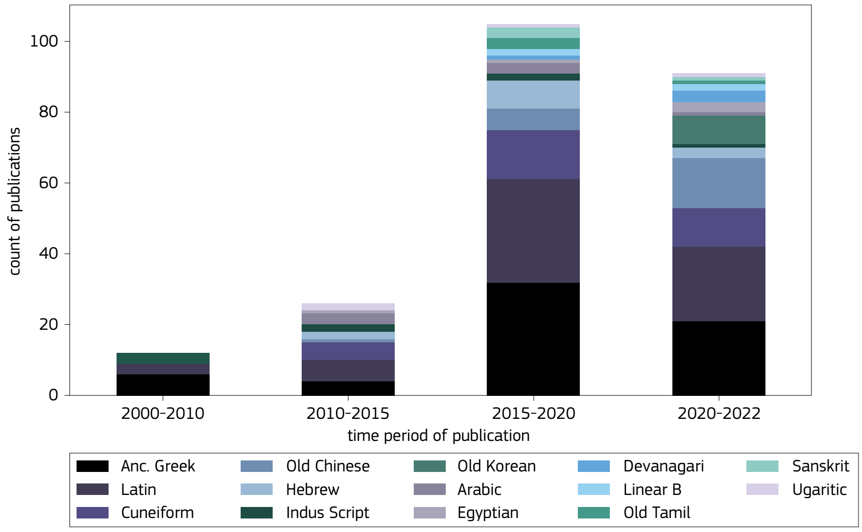
Virtual restoration of inscriptions includes the recovery of ancient texts, documented in inscriptions, allowing the present generations to gain insights on the evolution of writing but also the narratives of those cultures. However, due to age or external factors, ancient texts can be subjects of degradation and partial loss of information, which eventually leads to a loss of context for fields such as history, archaeology and literature.

Figure 32. Restoration of a damaged inscription, recording a decree from 485/4 BCE concerning the Acropolis of Athens (IG I3 4B, CC BY-SA 3.0, WikiMedia).

Source: Assael et al. (2022)

A survey published in the last few years (Sommerschield et al., 2023) on the prevalence of the uptake of ML in the study of ancient languages, as shown in Figure 33, found that there has been a significant increase in the interest in tackling tasks in this space, ranging from the determination of authorship to the restoration of a damaged or partially incomplete inscription.

Figure 33. Amount of literature contributions related to ML-based studies into ancient languages domain.



Source: JRC's own elaboration based on data from literature contribution (Sommerschield et al., 2023).

Focusing on inscription restoration, the first ancient text missing character restoration model was Pythia. Its task was to predict characters from a partial text input (Assael et al., 2022). On a character level, the error rate that was reported for Pythia was 30.1%, while human epigraphists were assigned a 57.3% error rate. Pythia was the result of the collaboration between researchers from DeepMind and the University of Oxford (Assael et al., 2019). Pythia was made available as a Python notebook interface and an offline version, available on GitHub. Another example of a model tackling the task of restoration of ancient Greek inscriptions was Ithaca (Assael et al., 2022). Crucially, Ithaca, using a transformer-based architecture, can be deployed to enhance and elevate the work of historians. When historians were assigned with the task of text restoration, their accuracy rose from 25% to 72% when not using and using Ithaca respectively. In terms of the other tasks Ithaca was trained on, namely geographic and chronological determination, on the former its accuracy was demonstrated to be 71% and on the latter, it managed to only deviate by 30 years from the ground-truth-assigned date (Assael et al., 2022). Ithaca was also made available as an interactive Python notebook and on GitHub.

In Asia, an example of DL to decode ancient Chinese inscriptions was proposed by Wang et al. (2025). These inscriptions could contain information on the Chinese economy, politics, culture, and way of life. The team built and trained a NN with a twofold objective: increasing the efficiency of restoration tasks and promoting the need for protection of items of cultural heritage. Their pipeline used a combination of Natural Language Processing (NLP) and Computer Vision (CV) and was shown to perform better than pipelines deploying only NLP or only CV (Z. Wang et al., 2025). Another instance of an ancient script being the object of an investigation into the deciphering of ancient text has been a study on artefacts from the Indus Valley Civilisation (IVC), Harappan Civilisation, dating around 2600 BCE to 1900 BCE. Dixit et al. (2025) proposed a pipeline which deploys three DL models to process and archive Indus Script and extract Motif information.

By uncovering information about the human past, societal ways of life and perceptions of the world can be understood, and this unearthed evidence is a testament to our historic continuity, either through the discovery of sites or through the interpretation of ancient written languages. The localisation of sites is key in ensuring their conservation, protection and restoration. Teams are already using AI to leverage the volume of information that resulted from the adoption of satellite and LiDAR technologies. Similarly, AI has been deployed to augment the task of restoring and attributing inscriptions across a breadth of ancient languages. So, it is clear that AI has and will shape how archaeological evidence is not only found but also incorporated to the existing body of knowledge.

3.3.1 DATA, MODELS AND INFRASTRUCTURE

In archaeology, AI is revolutionising research through advanced models, robust infrastructures, and comprehensive databases. AI technologies empower the analysis of virtual inscriptions, enhancing the accuracy and preservation of ancient texts. Additionally, AI-driven site detection uncovers potential archaeological sites, optimising exploration efforts. These advancements are supported by dynamic databases that store and integrate vast archaeological data, offering deeper insights into past societies.

Table 7. Collection of datasets related to the fields of site detection and inscription restoration.

Field	Dataset	Open source	Region	Reference
Site detection	Sentinel	Y	EU	Copernicus Sentinel Data[54]
	ASTER GDEM	Y	US	NASA's Earth Science Data and Information System (ESDIS)[55]
	Landsat-8	Y	US	NASA/USGS Landsat Science[56]
	OpenTopography	Y	US	(Krishnan et al., 2011)
	EpHEMERA	Y	EU	EpHEMERA Project[57]
	Odyssey Spatial Data Infrastructure (SDI)	Y	UK	(Sá et al., 2024)
	OpenHistoryMap	Y	EU	(Bernardoni et al., 2017)
	ArchaeoSTOR Map	Y	US	(Matsui et al., 2012)
Inscription restoration	PHI	Y	US	Packard Humanities Institute (PHI)[58]
	Chinese Inscription Rubbing Image (CIRI)	Y	US	(S. Zhu et al., 2024)
	Epigraphic Database Roma (EDR)	Y	EU	EAGLE's EDR Project[59]
	Asia Inscriptions Database	Y	UK	Asia Inscriptions Online (AIO)[60]
	Online Corpus of the Inscriptions from Ancient North Arabia (OCIANA)	Y	UK	Khalili Research Centre's OCIANA Project[61]
	Epigraphic Database of Ancient Asia Minor (EDAK[62])	Y	EU	University of Hamburg Open-Access Portal's EDAK Database[63]
Other	Archaeology Data Service (ADS)	Y	UK	ADS Organisation[64]
	The Digital Archaeological Record (tDAR)	Y	US	Arizona State University's Centre for Digital Antiquity Digital Repository[65]
	Data Station Archaeology	Y	EU	Dutch national centre of expertise and repository for research data (DANS)[66]

Source: JRC's own elaboration.

54 https://dataspace.copernicus.eu/data-collections/copernicus-sentinel-data/.

55 https://www.earthdata.nasa.gov/about/esdis/.

56 https://landsat.gsfc.nasa.gov/.

57 https://ephemera.cyi.ac.cy/about/.

58 https://packhum.org/.

59 http://www.edr-edr.it/.

60 https://siddham.network/about/.

61 https://krc.web.ox.ac.uk/article/ociana/.

62 Acronym from the original German resource name 'Epigraphische Datenbank zum Antiken Kleinasien'.

63 https://www.oa.uni-hamburg.de/datenbanken/epigraphik.html/.

64 https://archaeologydataservice.ac.uk/.

65 https://www.tdar.org/.

66 The Dutch national centre of expertise and repository for research data is called DANS, which stands for 'Data Archiving and Networked Services'. The dataset is available here: https://dans.knaw.nl/en/archaeology/.

Table 7 provides a summary of some datasets used in the archaeological field. It demonstrates the diverse range of data sources used, such as textual corpora, satellite imagery and LiDAR data, and the different geographical characteristics of the data sources. As pointed out by Gattiglia (2025), *training data in this field seems to be limited as well as standards for data combination, which contrast with previous use cases in the hard sciences where researchers work and produce substantial volumes of data*. The same author also argues that this is due to the existing ecosystem of small research teams and existing barriers to access such as copyright or poor metadata. In order to tackle the fragmentation of data, the EU funded the Integrating Activity Advanced Research Infrastructure for Archaeological Dataset Networking in Europe (ARIADNE).[67] Building on this, the ARIADNE project still supports and has expanded its mission to make archaeological data and services accessible to researchers. The ARIADNE Portal contains the data registry and a service which are employed to manage the datasets, collections, vocabularies, metadata schemas and mappings. As part of its services, the portal includes the option to access *compute infrastructure*. Through the ARIADNE Plus Lab,[68] the research community can process, visualise and analyse data from both the ARIADNE registry as well as their own. Furthermore, by accessing ARIADNE, landscape data can also be generated and processed, using open-source toolkits. These terrain datasets are designed for visualisation in real-time and web streaming.[69] Moreover, the ARIADNE dataset is comprised of data provided by researchers and organisations who wish to make datasets available. This project has more than 25 members[70] and the portal includes as of July 2025 more than 4 million records, covering a period from the earliest Hominids to the Cold War and spanned four continents. An example of infrastructure which supports the reusability of data is DANS,[71] the collaboration between the Royal Netherlands Academy of Arts and Sciences and Dutch Research Council, which functions as the national centre of expertise of the Netherlands and repository for research data. It facilitates researchers in making their data available for reuse, promoting the initiation of new research endeavours and enhancing the verifiability and reproducibility of published studies, with a collection exceeding 300,000 datasets.

[67] The Dutch national centre of expertise and repository for research data is called DANS, which stands for 'Data Archiving and Networked Services'. The dataset is available here: https://dans.knaw.nl/en/archaeology/.

[68] https://ariadne.d4science.org/web/ariadneplus_lab/.

[69] https://seth.itabc.cnr.it/services/landscape/.

[70] https://www.ariadne-research-infrastructure.eu/partners/.

[71] https://dans.knaw.nl/en/.

Table 8. A collection of examples of architectures reported to have been deployed in the fields of site detection and inscription restoration, along with their respective application area.

Field	AI Model and/or architecture	Application	Open source	Reference
Site detection	YOLOv3-based Multi-Scale Relief Model (MSRM) + Random Forest (RF)	Burial mound detection (Galicia, Spain)	Y	(Berganzo-Besga et al., 2021)
	RF-based model	Mound mapping (Cholistan desert, Pakistan)	Y	(Orengo et al., 2020)
	Random Forest	Burial mound detection (Carnac, France)	Partial	(Guyot et al., 2018)
	Vision Transformer	Burial mound detection (Alto Minho, Portugal)	N	(Canedo et al., 2024)
	Support Vector Machines (SVMs)	Stone-walled structures detection (Mpumalanga, South Africa)	Partial	(Mohlehli et al., 2023)
	Support Vector Machines (SVMs)	Structural remains detection (Chun Castle, UK)	Partial	(Kadhim and Abed, 2021)
	CarcassonNet (CNN)	Hollow roads detection (the Netherlands)	N	(Verschoof-van der Vaart and Landauer, 2021)
	U-Net (DL)	Ancient site detection (southern Mesopotamia)	Y	(Casini et al., 2023)
	ChatGPT-3.5-based experiments (prompt engineering)	Roman oil lamp typological investigation	N	(Lapp and Lapp, 2024)

Field	AI Model and/or architecture	Application	Open source	Reference
Inscription restoration	Pythia (sequence-to-sequence + LSTM)	Ancient Greek text decoding	Y	(Assael et al., 2019)
	Ithaca (Transformer)	Ancient Greek text decoding and chronological and geographical attribution	Y	(Assael et al., 2022)
	CIRoBERTa (NLP) + CISwin (CV)	Chinese character restoration	N/A	(Z. Wang et al., 2025)
	Ancient Script Recognition Network (ASR-net) + Motif Identification Network (MI-net)	IVC script and motifs extraction and identification	Y	(Dixit et al., 2025)

Source: JRC's own elaboration.

Table 8 provides a summary of a set of selected AI models and application tasks used in archaeological research. It demonstrates the variety of AI techniques employed, ranging from traditional ML algorithms to DL and LLMs, learning from diverse data modalities, including text, images or LiDAR data. Especially in the case of voluminous satellite or LiDAR data, dedicated infrastructure is necessary to host and visualise it, making it accessible to researchers. Along this axis, McCoy (2017) discussed the volume of geospatial data captured with satellite imagery and LiDAR can be challenging to handle. An open-access, 3D interactive online geo-database of architectural and archaeological heritage sites located in the Southeastern Mediterranean basin is the Online 3D Database System for Endangered architectural and archaeological Heritage in the south Eastern MEditerRAnean area (EpHEMERA), developed by the Cyprus Institute's Science and Technology in Archaeology and Culture Research Centre (STARC). This service allows users to visualise the structures, query the database and access geometric and morphological information. The database hosts structures from archaeological excavations, ancient buildings, archaeological areas. In order to support domain experts in handling Geospatial data, like LiDAR, a platform which integrated this data with a processing environment was created, the Odyssey Spatial Data Infrastructure (SDI) (Sá et al., 2024). The platform focuses on data concerning archaeological sites and offers functionalities to annotate, validate and visualise the spatial data. Through this work, it is possible to process remote sensing data and build ML algorithms. Other examples that fall under platforms that make archaeological geodata available on the web are OpenHistoryMap (Bernardoni et al., 2017) and ArchaeoSTOR map (Matsui et al., 2012), which is a testament to the need for such platforms in order to have readily-accessible tools and ways to visualise such high volume of information.

Albeit different tasks, both site detection and inscription restoration rely on the availability of extensive datasets and appropriate infrastructure for the processing of this data and the augmentation of the work of the research community with the deployment of AI models. In turn, teams investigating the potential of such models have proposed a variety of architectures, ranging from machine learning to Deep Neural Networks (DNNs). An examination of an array of examples across the two tasks revealed the common needs that emerge, in terms of compute and data storage infrastructure, and the role that platforms that allow the processing of complex, fragmented, multi-source and multi-modal data can play in making AI experimentation and adoption more accessible.

3.2.2 SAFE AND TRUSTWORTHY AI, ETHICAL CONSIDERATIONS AND CHALLENGES

Together, the presented studies exemplify how AI-enabled methods are expanding the interpretative capabilities of archaeology beyond the traditional methodologies. The integration of AI in archaeology, supported by the EU's Horizon Europe program, has the potential to revolutionise the field. However, as AI-driven discoveries become more prevalent, it is essential to prioritise ethical considerations, such as cultural sensitivity, heritage protection and community involvement (UNESCO, 2022). Through those recommendations, UNESCO urged member states to *incorporate AI systems, where appropriate, in the preservation, enrichment, understanding, promotion, management and accessibility of tangible, documentary and intangible cultural heritage* (UNESCO, 2022). At the same time, Horizon Europe emphasises the importance of *responsible and human-centric development and use of AI in science*, which can be applied to various sectors, including cultural heritage (European Commission: Directorate General for Research and Innovation, 2024). In this context, several projects have been funded by Horizon Europe program to ensure security and protection of cultural heritage, such as RITHMS[72] (Research, Intelligence and Technology for Heritage and Market Security), ANCHISE[73] (Applying New solutions for Cultural Heritage protection by Innovative, Scientific, social and economic Engagement), ENIGMA[74] (Endorsing safeguarding, protection, and provenance management of cultural heritage) and AURORA[75] (Artwork Unique RecognitiOn and tRacking through chemicAl encoded data, miniaturised devices and blockchain alliance).

Regarding data, Gattiglia (2025) discusses the human interpretation and action involved in the so-called 'travelling of data', i.e. from primary data (unprocessed data coming from the direct observation of the archaeological evidence) to secondary data (processed data made available for re-use) and to tertiary data (interpreted data derived from other researcher's analysis and reuse). In this respect, and in addition to concerns pertaining to the 'reductionist approach' of creating AI-ready data, the same contribution mentions that the creation of datasets is neither a neutral nor a purely technical process but incorporates political, cultural and social choices. These considerations are particularly relevant in disciplines with strong cultural grounds. The *CARE* and *FAIR* principles serve as vital frameworks for guiding the ethical and efficient management of data, addressing both cultural sensitivities and technical standards (Carroll et al., 2021):

- The FAIR (Findable, Accessible, Interoperable and Reusable) principles are a set of guidelines designed to improve the management and sharing of data. These criteria aim to enhance data transparency and usability, making it easier for researchers and organisations to access, integrate and apply data across various fields and disciplines (Wilkinson et al., 2016). By adhering to FAIR, data are effectively managed and shared, promoting a more open and collaborative scientific environment.

- The CARE (Collective benefit, Authority to control, Responsibility and Ethics) principles for Indigenous data governance ensure that Indigenous data are managed respectfully, aligning with Indigenous rights and perspectives (Carroll et al., 2020). By adopting these guidelines, institutions and researchers can foster respectful and equitable partnerships with Indigenous communities.

As a result, despite recent advances in AI uptake in this field, broader integration of AI into archaeology remains limited. Bellat et al. (2025) reviewed ML studies in archaeology published between 1997 and 2022 (excluding LLMs) and found that the field is dominated by supervised classification models, particularly artificial NNs and ensemble learning techniques, which together account for nearly two-thirds of applications. The most common use cases involve automatic structure detection and artefact classification, while more interpretive applications such as taphonomic analysis or predictive modelling remain rare. Moreover, the authors conclude that a major limitation lies in the narrow focus of many ML applications. Rather than advancing theory-driven interpretation, most studies concentrate on data processing tasks, revealing

72 https://rithms.eu/.

73 https://www.anchise.eu/.

74 https://eu-enigma.eu/.

75 https://www.aurora-euproject.eu/.

a persistent gap between computational AI methods and archaeological thinking. Bridging this gap is essential to fully integrate AI into interpretive archaeological frameworks.

Further challenges in this domain have been identified regarding the adoption of AI and the connected concept of trustworthy AI (Gattiglia, 2025). About the former, concerns mainly relate to data, in particular their availability, fragmentation and large volume, which introduce complications when attempting to combine sources of information. Moreover, there is an inherent complexity in the artefacts and sites themselves due to their patterns and the influence of external factors such as human actions and natural processes. Overcoming biases in the dataset and assigning chronological and geographical attributes also hinder the data processing process. Concerning the latter challenge about AI in archaeology, transparency constitutes a crucial ethical aspect. A comprehensive understanding of data distortions and the operational principles of AI algorithms is essential for archaeologists to establish trust in a system's outputs and integrate the results into their research studies. The application of AI models in the interpretation of archaeological findings underscores the importance of explainability and transparency, particularly with regard to potential biases inherent in the underlying data. Failure to address these concerns may result in the perpetuation of biases and the solidification of flawed interpretations as historical facts, thereby undermining efforts to redefine the field and challenge existing power dynamics. Notably, a recent study (Montgomery and Fryer, 2023) has sought to critically examine and challenge dominant power structures in archaeology, with a specific focus on collaborative research methodologies informed by decolonisation and Black feminist perspectives.

3.3.3 SKILLS AND INNOVATION

Although the use of AI is driven by the analysis of different data sources and time scales, they all rely on integrating qualitative and quantitative methods. Archaeology, in particular, exemplifies the scientific value of *interdisciplinary* synthesis. According to Gattiglia (2025), the successful uptake of AI in archaeology requires interdisciplinarity collaboration between computer scientists and archaeologists to bridge the knowledge gap and develop robust tools and methodologies. As discussed, there are specific web servers dedicated to improving the accessibility and processing of archaeological big data. Along with this, the ability to understand the fundamentals of AI models and be in a position of using their capabilities is becoming of increasing importance. Yet, at the same time, *the cultivation of a high level of scientific domain expertise is critical*. Thus, the 'archaeologist of tomorrow' has a strong scientific knowledge, while at the same time can navigate tools that can accelerate their workflow, like web servers, databases and AI models. Representative studies suggest how the adoption of an AI model enhances the work of a scientist in inscription attribution (Assael et al., 2022). This combination of disciplines is not novel in archaeology, since it has long constituted a convergence point of scientists with backgrounds in physics, chemistry, biology, computer science and the humanities to reconstruct the past from incomplete and diverse evidence. With the accelerating increase in the production of data and types of data that are able to be captured, the ability for efficient *archival search* remains fundamental and relates to both the ability to retrieve information from physical and digital sources.

A demonstration of how vast the shift is in regard to skills for a future archaeologist is the breadth of innovations that have already been implemented in the field. Examples of those changes are data management systems that enable a Geographic Information System (GIS) to be accessed via a mobile phone, the introduction of robotics and the deployment of AI across multiple tasks along with existing methodologies. Virtual Reality (VR) and Augmented Reality (AR) are being explored in order to contextualise the discoveries and leverage this technology to support educational ventures. Combining two of the afore-mentioned innovations is the 'Reconstructing the Past: Artificial Intelligence and Robotics meet Cultural

Heritage' (RePAIR) project,[76] which is driven by the Ca' Foscari University of Venice, the Ben-Gurion University of the Negev of Israel and the Italian Institute of Technology and supported partially by EU funds. The aim of the project was to deploy robotics, CV and AI modes for cases of a high degree of fragmentation, which would have made the restoration unfeasible for operators, despite the availability of specialised software. As discussed, Google DeepMind has been involved, in collaboration with academia, in the development of Ithaca, to restore and attribute ancient Greek inscriptions (Assael et al., 2022). These examples highlight the achievements as well as the possibilities for accelerating innovation within the field of archaeology, driven by the whole ecosystem of academia, large organisations and startups and how the synergies observed are in line with the projection of the future-ready archaeologist, which rests on deep expertise and interdisciplinarity.

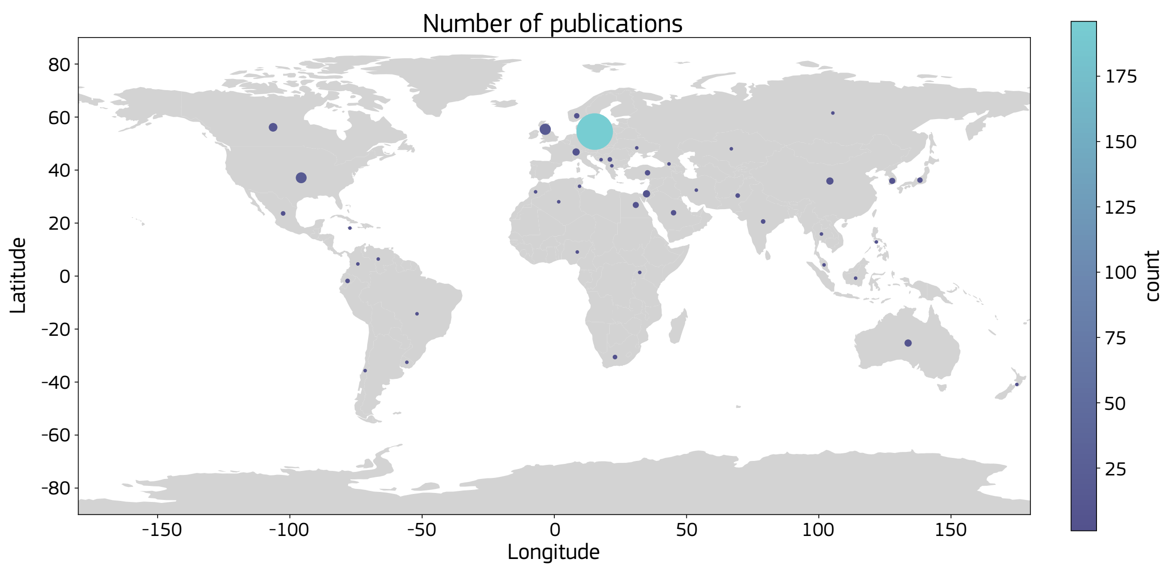
3.3.4 RESEARCH COMMUNITY

The application of AI to ancient site discovery and virtual inscription restoration is emerging as a niche but highly interdisciplinary field. It brings together archaeology, epigraphy, computational linguistics, remote sensing and ML, with the goal of enhancing the documentation, interpretation and preservation of cultural heritage. Progress in this domain has been made possible by the convergence of digitised corpora, imaging technologies and AI techniques, as well as by collaborative efforts spanning humanities research institutions, computer science labs and public heritage infrastructures. To explore the development of this research community, 315 publications indexed in Scopus between January 2020 and May 2025 (retrieved following the methodology described in Section 1.3) are examined by filtering for articles, reviews and conference papers containing the keywords the keywords described in Table 1. This analysis provides a quantitative snapshot of the field's geographical distribution (Figure 34), institutional diversity (Figure 35 and Figure 36) and collaborative patterns across regions and sectors (Figure 37).

In terms of geographical distribution (Figure 34), research in this area is strongly concentrated in the EU, which accounts for 196 publications (62.2%). Other contributing regions include the UK (16; 5.1%), the US (15; 4.8%), Switzerland (8; 2.5%), and China (6; 1.9%). This distribution reflects the EU's leading investment in cultural heritage research infrastructures, as well as its long-standing academic networks in digital archaeology and epigraphy.

[76] https://www.repairproject.eu/.

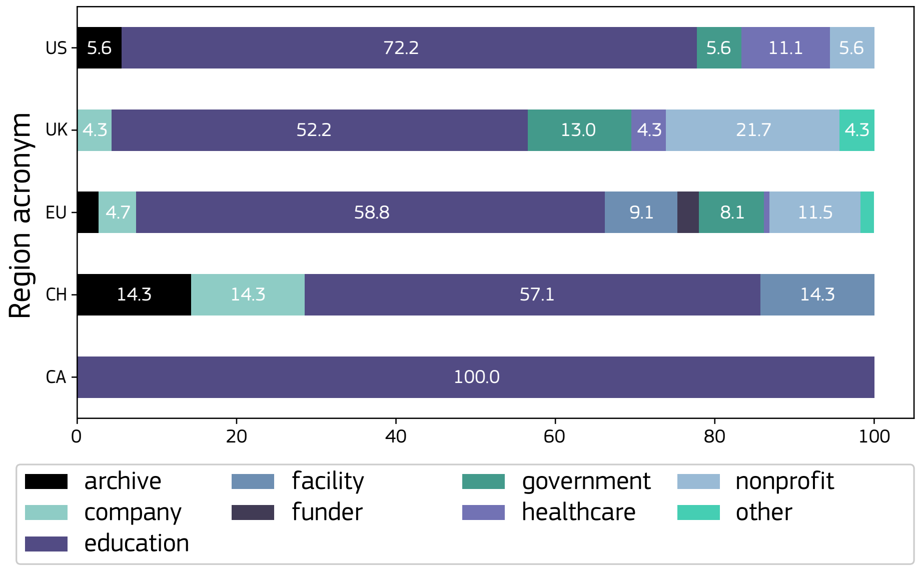
Figure 34. Global geographical distributions of scientific literature contributions related to 'ancient site discovery' and 'virtual inscriptions restoration' topics, indexed by Scopus between Jan. 2020 and May 2025.



Source: JRC's own elaboration based on divinAI methodology.

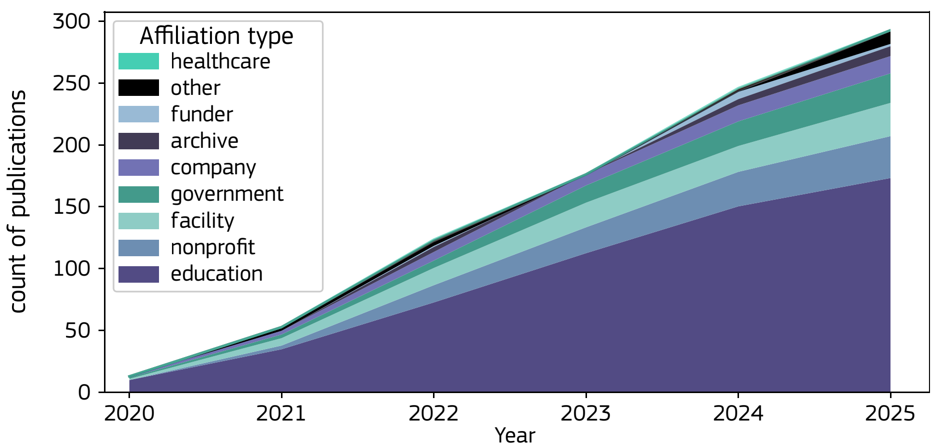
At the global level, research in this field is largely led by educational and academic institutions, which account for the vast majority of contributions (Figure 35). A more detailed analysis of the European landscape confirms this trend: within the EU, most publications originate from universities, followed by non-profit organisations, public research facilities, government bodies, and less prominently, private companies (Figure 36). This distribution reflects the strong anchoring of the field within the academic and cultural heritage sectors. The availability of digitised archives, shared infrastructure and open-source tools has enabled widespread participation, even among institutions with limited access to advanced computational resources.

Figure 35. Distribution of 'ancient site discovery' and 'virtual inscriptions restoration' literature contributions by type of establishments and geographical area. Considered contributions are indexed by Scopus between Jan. 2020 and May 2025.



Source: JRC's own elaboration based on divinAI methodology.

Figure 36. Amount of 'ancient site discovery' and 'virtual inscriptions restoration' literature contributions across EU countries by type of establishments over the years from Jan. 2020 to May 2025. Considered contributions are indexed by Scopus.



Source: JRC's own elaboration based on divinAI methodology.

Beyond institutional diversity, collaboration networks also shape the research landscape in this domain. Due to the strong concentration of activity in the EU, cross-regional collaborations are relatively modest in scale but most commonly involve partnerships with the UK, the US, Switzerland and Canada (Figure 37(a)). At the institutional level within the EU, the most frequent collaborations occur between educational institutions and non-profit organisations, as well as between universities and government bodies (Figure 37(b)). These patterns reflect the academic anchoring of the field and its strong ties to the public and cultural heritage sectors, with comparatively limited engagement from commercial actors.

Figure 37. Collaboration patterns in 'ancient site discovery' and 'virtual inscriptions restoration' fields. Amount of co-authored literature contributions by (a) geographical area and (b) type of establishments (only EU). Considered contributions are indexed by Scopus between Jan. 2020 and May 2025.

Source: JRC's own elaboration based on divinAI methodology.

Ancient site discovery and virtual restoration of inscriptions: Take-home messages

- AI-driven methodologies, such as the analysis of remote sensing data (LiDAR and satellite imagery) and the application of DNNs, are significantly enhancing the discovery and preservation of archaeological sites and cultural heritage.
- In the field of epigraphy, AI models have demonstrated the ability to surpass human performance in tasks such as restoring partially damaged ancient inscriptions, providing a powerful new tool for historical research.
- This is a highly interdisciplinary domain, with a research community strongly anchored in EU and supported by initiatives like ARIADNE, which focuses on federating fragmented data and making it more accessible to researchers.
- The primary challenges involve overcoming data fragmentation and ensuring ethical considerations, particularly those related to cultural sensitivity and bias, are at the forefront of AI deployment.

4.

CHALLENGES AND OPPORTUNITIES

The assessment of the use of Artificial Intelligence (AI) in science reveals that AI is already a widely used technology by researchers in varied fields, tasks, disciplines and steps of the scientific process. As an example, AlphaFold, developed in the biology domain, and specifically in the protein structure prediction field, showed the potential of AI for the acceleration of drug discovery, the enhancement of our understanding of disease mechanisms, and the unlock of innovative biotechnological applications. The use of AI for scientific discovery is both an opportunity to drive a new wave of innovation and a complex policy challenge. Creating an enabling environment for AI in science requires coordinated action across funding frameworks, governance mechanisms, infrastructure development and talent strategies. These policy domains are interconnected, and policy development is taking place in an environment where technological capabilities and research practices are shifting (European Commission: Directorate General for Research and Innovation and Montgomery, 2025). National policies for AI in science must address this complexity, as different patterns of adoption reflect a range of technical, institutional, and cultural influences (Bianchini et al., 2024).

In this section, the key opportunities and challenges associated with the integration of AI in scientific research are highlighted, and the relevant ingredients necessary for a responsible and effective uptake of AI technologies are discussed, including the need for critical evaluation, transparency and accountability.

4.1 Data, models and infrastructure

The access to high-quality data, computing infrastructure and scientific models have been found as the basic building blocks or facilitators for the uptake of AI in science.

Scientific data is the main ingredient for the development of AI models for science. Publicly available databases, such as the Protein Data Bank, discussed in Section 3.1, supported by the European Molecular Biology Laboratory (EMBL), have been instrumental in the development of cutting-edge models in the field. This dataset exemplifies the strength of community-driven and international collaboration for the building of high-impact datasets. However, AlphaFold and other protein structure prediction methods suppose a big leap in accuracy but are far from being perfect, due to constraints in the training (experimental) data. As is well-supported by the literature, the generation of scientific training data and standards for data combination appears to be more limited in certain fields of the social sciences and computational humanities, such as archaeology, in contrast with different research areas in the hard sciences. As is well-supported by the literature, the generation of scientific training data and standards for data combination appears to be more limited in certain fields of the social sciences and computational humanities, such as archaeology, in contrast with use cases in the hard sciences. This disparity stems from the distinct nature of their core challenges: while hard sciences face obstacles related to the physical world (Tang et al., 2025), hardware (Desai et al., 2025; Z. Liu et al., 2024) and the need for explainable causal models (Makke and Chawla, 2024), the social sciences and humanities confront more fundamental issues of inherent subjectivity (Bail, 2024), cultural bias (Binz et al., 2025) and the nuanced, often unquantifiable nature of their data (Xu et al., 2024). Consequently, the ability of AI to generate truly representative training data or to set universal standards for combining qualitative, context-dependent information is significantly more constrained in these fields.

The training of scientific models from this data often demands access to High-Performance Computing (HPC) clusters and results in a high resource cost, which is sometimes only accessible to large laboratories. In this respect, community resources and shared infrastructures facilitate access and uptake even from European laboratories or research fields that may have small local computing budgets, as seen in all analysed fields. There is as well research being done to reduce model size leading to decrease in infrastructure needs and related financial and environmental cost. European Union (EU) Initiatives such as the *AI Factories*[77] can act as catalysts in the realisation of large-scale AI projects in different scientific fields.

A key emerging trend is the development of scientific foundation models. These are large, General-Purpose AI (GPAI) models pre-trained on vast, often multimodal, scientific datasets. Unlike traditional models tailored for a single task, foundation models are designed to learn fundamental relationships and principles across a broad scientific domain (e.g., molecular biology, materials chemistry, or geospatial data) (Wang et al., 2023). This allows them to perform zero-shot learning, i.e. solving new problems without specific training, and facilitates transfer learning, where a pre-trained model can be rapidly adapted for a specialised task. For instance, a model already trained on a massive corpus of molecular structures and properties could be fine-tuned to predict a novel material's activity or a protein's drug-binding capability. While these models promise to accelerate discovery, they also bring new challenges. The immense resources required for their development concentrate power in a few large institutions, potentially creating dependencies for the wider scientific community (Binz et al., 2025). Ensuring these models are developed under open science principles, with shared data, weights and transparent methodologies, is therefore critical for maintaining an equitable and innovative research landscape.

Regarding the *development* of scientific AI models, *a varied set of AI techniques and approaches are found in the literature*, from the most traditional to novel developments in Machine Learning (ML), so not being restricted to recent developments in foundation and Large Language Models (LLMs). For instance, our analysis in Section 3.3 revealed the use of different ML architectures, from random forests to Deep Neural Networks (DNNs), used in the archaeological field. Generative AI (GenAI) models are the most promising ones in the field of material discovery (see Section 3.2) based on Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs) or autoregressive transformers and diffusion models, and their size is relatively moderate (i.e. five or six orders of magnitude) compared to state-of-the-art LLMs. However, these models are instrumental in several scientific tasks, notably in literature review or scientific writing, as discussed in Section 2. In addition, ML techniques, when used in specific scientific tasks, evolve and incorporate specific and realistic contexts from that particular task. This is the case for instance in automated models used for protein design which incorporate molecular context in which proteins function, as presented in Section 3.1.

In terms of *model sharing*, it is observed that scientific AI models are frequently released as web tools, which lowers the infrastructure barrier to entry for the use (not training) of these models. In this context, and connected to the concept of *open science*, *the uptake of AI brings the need to incorporate concepts linked to open infrastructures, data and open-source AI as facilitators for research reproducibility and to lower the accessibility barrier*. For instance, as commented in Section 3.1, the protein structure prediction field has in its heart the open science principles, with models being trained on publicly available data and in turn released as open-source frameworks.

A disparity in resources exists between larger and smaller countries, with many scientists in under-resourced EU nations facing significant barriers to access HPC resources and expertise ((European Commission: Directorate General for Research and Innovation, Group of Chief Scientific Advisors and King, 2025; European Commission: Directorate General for Research and Innovation and Groznik, 2025). This is exacerbated by the fact that few countries have dedicated HPC facilities for AI, with many still focused on traditional applications, which can hinder AI research (European Commission: Directorate General for Research and Innovation, Group of Chief Scientific Advisors and King, 2025). Industry dominance in computing

[77] https://digital-strategy.ec.europa.eu/en/policies/ai-factories/.

resources also makes it easier for private companies to attract talent and invest in research, further widening the gap and raising concerns about equitable access (Ahmed et al., 2023). The immense resources required for the development of scientific foundation models concentrate power in a few large institutions, potentially creating dependencies for the wider scientific community (Binz et al., 2025).

4.2 Skills and innovation

The uptake of AI in science generates new requirements for specialised expertise among researchers and research teams. The analysis on *skills* in the three selected deep dives (see Section 3) agree on *the need for hybrid (inter- and multidisciplinary) teams that combine expertise in engineering and computer sciences with specific domain expertise, where full autonomy does not have to be a goal in itself*. This ensures the factuality and soundness of scientific findings, and the tailoring of approaches to specific scientific data and research questions. The future-ready research in protein structure prediction, for instance, is proficient in both structural biology and AI/HPC principles, and the collaboration across establishments such as the one between DeepMind and EMBL has been shown to be crucial for AI innovation. The example of the archaeology field, and similarities in other social sciences domains, brings together disciplines traditionally far in terms of methodologies and approaches, making inter-disciplinarity fundamental to the fields' ability to generate meaningful scientific insights.

However, talent strategies across EU face significant challenges, including a lack of flexible career progression pathways and difficulties in retaining talent in research institutions due to competition with the private sector (European Commission: Directorate General for Research and Innovation and Groznik, 2025). The skills needed for AI in science span a wide range, from technical AI capabilities to interdisciplinary collaboration and data management. Without clear career directions and adequate funding, the public sector struggles to compete with the lucrative opportunities offered by big tech companies (European Commission, Directorate General for Research and Innovation and Groznik, 2025; Jurowetzki et al., 2025). This challenge is particularly acute for 'hybrid' roles that combine expertise in AI with deep scientific domain-specific knowledge.

The EU has also launched several initiatives aimed at fostering interdisciplinary collaboration and digitally-enabled research in the arts and humanities. For instance, the Digital Research Infrastructure for the Arts and Humanities (DARIAH)[78] brings together scholars from diverse disciplines to support innovative research and teaching methodologies across the arts and humanities, covering fields such as the intersection of AI and music, as well as the analysis and linking of bibliographical data. To facilitate the effective implementation of AI in various scientific domains, a range of educational resources, toolkits, encyclopaedias, standards and best practices have been developed, providing a comprehensive set of tools and guidelines for researchers and educators to efficiently leverage AI in their daily work activities. In this context, the EC's AI Office has taken several actions to promote AI literacy. It has notably published a set of FAQs to assist stakeholders, hosted webinars and created a *living repository*[79] of best practices to help organisations tailor training to their staff. The AI Office is also rolling out broader initiatives like the 'AI SkillsAcademy' under the **AI Continent Action Plan**.[80]

4.3 Ethical and legal considerations

The use of AI in scientific research raises a variety of ethical and legal concerns that must be addressed to ensure responsible development and deployment. These concerns span issues such as privacy and data protection, bias and fairness, transparency and explainability, human oversight, and the broader impact on environmental and

78 https://www.dariah.eu/.

79 https://digital-strategy.ec.europa.eu/en/library/living-repository-foster-learning-and-exchange-ai-literacy.

80 https://digital-strategy.ec.europa.eu/en/factpages/ai-continent-action-plan.

societal well-being. These topics are closely aligned with the seven key requirements defined by the 'Ethics Guidelines for Trustworthy AI' (European Commission, Directorate-General for Communications Networks, Content and Technology and High-Level Expert Group on Artificial Intelligence, 2019).

The European Research Area (ERA) Forum, a key stakeholder body in the European research landscape, has also highlighted several key principles for the responsible use of GenAI in research in their 'Living guidelines on the responsible use of generative AI in research' (European Commission, Directorate General for Research and Innovation, 2025). These principles are built on the 'European Code of Conduct for Research Integrity' (ALLEA, 2023) and the already mentioned trustworthy AI guidelines, and emphasise reliability, honesty, respect and accountability. Acknowledging the systemic pressures within the contemporary research landscape, these guidelines affirm the indispensable role of human oversight. AI systems are to be regarded as instruments in the research process, not as authors or co-authors. Consequently, the researcher is directly accountable for the validity and integrity of their scientific output, a duty that includes understanding and mitigating the tool's limitations, such as prompt bias (sycophantic behaviour) and the generation of erroneous citations. This notion of human responsibility intersects with several already well-established research practices. For example, researchers are accustomed to rigorous ethical review procedures and data protection regulations. However, the integration of AI requires extending these frameworks to account for the unique properties of algorithmic systems, such as their opacity, dependence on training data and probabilistic outputs.

Transparency is a core requirement. Researchers must use AI tools transparently and disclose their use in the research process, particularly when an AI tool has a substantial impact on the results. This includes being mindful of the *stochastic nature* of GenAI, which can produce different outputs from the same input. Confidence scores, uncertainty estimates and benchmarks are already used in certain fields (e.g. protein structure prediction; see Section 3.1) to support the interpretability and reproducibility of AI-assisted results. Another critical issue is AI interpretability, that is, the ability for both researchers and external observers to understand how an AI system arrived at a specific output. A lack of interpretability, particularly in black-box models like NNs, can hinder scientific trust, limit peer scrutiny and challenge reproducibility. This is vital not only for scientific accountability but also for interdisciplinary collaboration. Section 3.3 provides an example where interpretability helps bridge computational and archaeological thinking.

Bias and fairness are also prominent concerns. In scientific research, algorithmic bias may emerge from skewed training data, inappropriate model assumptions or cultural and contextual insensitivity (Purificato et al., 2024). In fields such as archaeology, the geographical origin or cultural background embedded in datasets may lead to distorted outputs if not properly accounted for. Moreover, AI systems can inadvertently reinforce existing scientific hierarchies, marginalise underrepresented methodologies or geographies, and limit epistemic diversity. Bias and fairness are also closely linked to the principle of *respect* for colleagues, research participants, society, ecosystems, cultural heritage and the environment.

Data governance, including privacy and intellectual property aspects, are also of paramount importance. Researchers are responsible for ensuring they have a clear legal basis for processing any personal data, in line with EU data protection rules, such as the General Data Protection Regulation (GDPR) (Regulation (EU) 2016/679). In high-stakes domains, AI can pose dual-use risks, where models developed for scientific inquiry may be repurposed for harmful or unintended applications. For example, advanced protein structure models could facilitate protein design applications with potential biosafety risks (see Section 3.1). While organisations have begun to assess and mitigate such risks, the fast pace of AI development makes it difficult to anticipate all possible downstream consequences.

From a regulatory perspective, the EU AI Act (Regulation (EU) 2024/1689) introduces a new legal framework for AI in Europe. While the act includes a 'research exemption' (Art. 6) that excludes scientific research from most obligations, its provisions may still apply once research outputs transition into deployable or marketable systems. This raises important questions about legal boundaries, responsibilities and oversight. As a result, researchers should be supported early on in considering compliance-by-design and ethics-by-design approaches, to align their AI systems with evolving regulatory frameworks and ethical norms. These considerations suggest that the ethical and legal dimensions of AI in science are not static or external constraints but active components of scientific quality, credibility, and societal responsibility. In this context, it is essential to acknowledge the potential risks associated with the uncritical adoption of AI tools, including the possibility of inventing references or identifying spurious correlations, as highlighted by recent studies (Buriak et al., 2023), which can have significant consequences for the research evaluation process and the integrity of scientific research (Eriksson et al., 2025).

4.4 Social dimension of science

Beyond legal and ethical considerations, the integration of AI in scientific research brings profound societal implications. AI is not only transforming how research is conducted but also reshaping social structures, cultural practices and public trust in science. This interplay has been described in the literature as creating 'algorithmically infused societies' (Wagner et al., 2021), where algorithmic systems act as bridges between science and society, influencing how information flows and decisions are made. In this context, Computational Social Science (CSS) has emerged as a key interdisciplinary field at the intersection of AI and the social sciences. CSS involves the application of AI and computational methods to study, simulate and understand complex social phenomena. It integrates tools such as ML, Natural Language Processing (NLP) and network analysis into disciplines including sociology, psychology, economics, political science and cognitive science (Conte et al., 2012; Lazer et al., 2009; Xu et al., 2024). This section explores the **dual perspective** of AI as both a tool for understanding society and a social force in itself.

- *Psychology and cognitive sciences*. LLMs have been used to detect depressive symptoms from social media data (K. Yang et al., 2023), infer personality traits from digital traces (Rao et al., 2023), and predict short-term mental states using smartphone and sensor data (Webb et al., 2025). In online learning, AI has been used to create adaptive learning systems that personalise educational content and pace for students (Bühler et al., 2025). While these tools aim to improve learning outcomes, a societal implication is the risk of entrenching educational inequality if the models are trained on biased data, perpetuating existing performance gaps.

- *Political science*. LLMs have supported media bias detection and political stance classification, which can be a valuable tool for researchers (Ziems et al., 2024). However, the use of such models also has a societal implication: if these AI tools are used to curate news feeds of flag 'biased' content without transparency, they could inadvertently shape public opinion and political discourse, creating a new form of algorithmic influence (Chioma and Lepe, 2024; Gandini et al., 2025).

- *Sociology*. AI has enhanced hate speech and misinformation detection. A relevant study showed that LLMs can identify implicit hate but may mislead users in borderline cases (Huang et al., 2023), while another combined framing theory with neural networks to detect reframed narratives in news content (G. Wang et al., 2025).

- *Finance*. Domain-specific LLMs trained on financial corpora have demonstrated strong performance on tasks like sentiment analysis, classification and information extraction in financial documents (S. Wu et al., 2023).

- *Linguistics*. AI models are now powerful tools for analysing vast linguistic corpora, allowing researchers to study historical language evolution (Liang et al., 2023) or trace the spread of misinformation in new ways (Saeidnia et al., 2025). However, these same models, when used as translation or communication tools, can homogenise language and cultural nuances, potentially leading to a loss of linguistic diversity over time (Bella et al., 2024). This demonstrates how a tool for scientific analysis can also have a profound societal impact on cultural heritage.

These examples underscore the dual role of AI as both a tool for understanding society and a social force in itself. As LLMs and other AI tools are integrated into the practice of social science, they bring with them opportunities for new forms of analysis, but also risks of reinforcing biases, reducing transparency and limiting contextual understanding (T. Hu et al., 2024)particularly the tendency to favor one's own group (ingroup solidarity. A critical societal implication of AI in science is thus the need for multidisciplinary oversight and collaboration. Technical experts, social scientists, and ethicists must work together to ensure that AI models and infrastructures reflect human values, safeguard social equity, and promote knowledge that benefits all members of society. In doing so, scientific AI development becomes not only a technical challenge but a societal effort.

4.5 Collaborations

AI's role in scientific research is driving new forms of collaboration. This is largely due to the need for interdisciplinary expertise, where technical knowledge of AI must be combined with deep domain-specific understanding to produce meaningful and reliable scientific outcomes. The central role of academia in advancing the field is clear, with a notable and growing contribution from external partners, including industry stakeholders such as start-ups and scale-ups, as well as public research institutions. For instance, a 325% increase in publications with private company affiliations from 2020 to 2024 is observed in the field of protein structure prediction (see Section 3.1), along with an ecosystem of startups specifically focused on leveraging advanced AI tools to enhance materials discovery research (see Section 3.2). This reflects the fact that industry is catching up on the use of AI in science, especially in recent years.

This *interdisciplinarity*, offering significant opportunities in the AI in science field, should be then embraced by research organisations, funding agencies or scientific association for the promotion of hybrid teams, evaluation panels, projects and communities across relevant disciplines. However, there are also some associated challenges. The report's deep dives (Section 3) highlight the need to bridge methodological and cultural divides between disciplines that have traditionally operated in silos. Technical experts and domain specialists must overcome differences in language, values and epistemic assumptions to co-create research that is both computationally sound and contextually relevant. As discussed in Section 4.4, AI, if not carefully governed, can reinforce existing scientific hierarchies or limit epistemic diversity by favouring data-rich, well-established areas of inquiry. This makes collaboration a vital tool for ensuring that AI-driven science remains inclusive, rigorous, and reflective of a diverse range of perspectives.

The investigation conducted in this report also reveals the *critical role of collaboration networks*, involving educational institutions, government bodies, research facilities, non-profit organisations and private companies, as well as the international dimension of scientific research involving AI. Even though collaboration networks might vary according to the discipline, EU-affiliated researchers in the studied fields were found to have a diverse array of collaborations across establishment and borders, compared to the other regions considered. Finally, *international cooperation* is recognised as a pivotal component of scientific progress, given that science is a global endeavour that transcends national boundaries and relies on the establishment of research networks and collaborations across the world.

5.

CONCLUSIONS

This final section synthesises the science-for-policy report's findings and interprets them in the context of the policy questions and objectives presented at its outset. The study confirms that Artificial Intelligence (AI) is a transformative, interdisciplinary and international force, already integrated into nearly all scientific disciplines. While AI offers immense opportunities to accelerate discovery and enhance research, its responsible uptake requires a coordinated and evidence-based approach to address significant technical, interdisciplinary and ethical challenges. The extensive analysis conducted across the scientific process and in three distinct deep dives, i.e. protein structure prediction, material discovery and computational humanities (specifically ancient site discovery and virtual restoration of inscriptions) reveals several key findings with potential policy implications supporting the adoption of the European Strategy for AI in Science.

First, the potential of AI is intrinsically linked to the principles of **open science**. The report demonstrates that publicly available data repositories, open-source models and collaborative infrastructures, such as the Protein Data Bank (PDB) and the materials databases discussed, have been instrumental in driving breakthroughs. Moreover, AI models have proven highly effective at **automated data processing** and **pattern recognition**, enabling researchers to efficiently analyse large multimodal datasets far beyond human capability, as seen in fields from genomics to astronomy. The report also notes the emerging trend of AI serving as a 'co-scientist' in **hypothesis generation**, accelerating literature analysis and suggesting novel ideas. Therefore, to sustain this pace of innovation and ensure reproducibility, policy actions should prioritise expanding and supporting these open ecosystems.

Second, the study highlights a growing demand for **robust computational and data infrastructure**. Advanced AI models, like the examined fields of protein structure prediction and material discovery, require significant resources for training and deployment. This need for High-Performance Computing (HPC) presents a strategic challenge and an opportunity for the EU to secure its leadership position in AI research. Targeted investment in these areas is essential to prevent a widening resource gap between large institutions and smaller research teams across the EU.

Third, the integration of AI necessitates the cultivation of a **multidisciplinary talent base**. As demonstrated by the report's deep dives, the most impactful research occurs at the intersection of AI expertise and deep domain-specific knowledge. The 'future-ready researcher' or 'future-ready research teams' is not only proficient in a scientific discipline but also skilled in data science and software engineering. This combination is crucial for effectively leveraging AI in tasks like **experiment design and optimisation**, where AI systems can accelerate and automate hypothesis testing in 'self-driving labs'. Policy should address the challenges of retaining such 'hybrid' talent in the public sector by offering competitive career paths and fostering interdisciplinary training programs.

Finally, while AI is widely recognised as a powerful tool, it is not without its risks. The report identifies concerns such as algorithmic bias, the potential for AI models to generate fabricated data (called 'hallucinations') and the risk of reinforcing existing paradigms, leading to the so-called 'epistemic drift'. These ethical considerations must be at the forefront of AI governance. This is particularly relevant in biology, where AI models for protein structure prediction are now used in lab settings, but their adoption in clinical practice remains limited due to the need for models that can accurately account for real-life conditions. The identified societal risks are also evident in the **communication of results**, where AI-assisted writing tools, while efficient, raise critical questions about authorship, originality and the trustworthiness of generated content.

REFERENCES

'Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance)', 2016.

'Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA relevance)', 2024.

Abdel-Rehim, A., Zenil, H., Orhobor, O., Fisher, M., Collins, R.J. et al., 'Scientific Hypothesis Generation by a Large Language Model: Laboratory Validation in Breast Cancer Treatment', arXiv, 2025, https://doi.org/10.48550/arXiv.2405.12258.

Abdul-Rahman, G., Zwitter, A. and Haleem, N., 'A Systematic Literature Review on the Role of Artificial Intelligence in Citizen Science', *Discover Artificial Intelligence*, No. 5, 2025, p. 165, https://doi.org/10.1007/s44163-025-00437-z.

Abeliuk, A., Gaete, V. and Bro, N., 'Fairness in LLM-Generated Surveys', arXiv, 2025, https://doi.org/10.48550/arXiv.2501.15351.

Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T. et al., 'Accurate Structure Prediction of Biomolecular Interactions with AlphaFold 3', *Nature*, No. 630, Nature Publishing Group, 2024, pp. 493–500, https://doi.org/10.1038/s41586-024-07487-w.

Afflerbach, B., Jiang, R., Tappan, J. and Morgan, D., 'Machine Learning Lab Module', *nanoHUB*, 2020, https://doi.org/10.21981/CPNK-XE48.

Afonja, T., Sheth, I., Binkyte, R., Hanif, W., Ambast, S. et al., 'LLM4GRN: Discovering Causal Gene Regulatory Networks with LLMs – Evaluation through Synthetic Data Generation', in: *ICLR 2025 Workshop on Machine Learning for Genomics Explorations*, 2025.

Agarwal, S., Sahu, G., Puri, A., Laradji, I.H., Dvijotham, K.D. et al., 'LitLLM: A Toolkit for Scientific Literature Review', arXiv, 2025, https://doi.org/10.48550/arXiv.2402.01788.

Agarwal, V. and McShan, A.C., 'The Power and Pitfalls of AlphaFold2 for Structure Prediction beyond Rigid Globular Proteins', *Nature Chemical Biology*, No. 20, Nature Publishing Group, 2024, pp. 950–959, https://doi.org/10.1038/s41589-024-01638-w.

Ahdritz, G., Bouatta, N., Floristean, C., Kadyan, S., Xia, Q. et al., 'OpenFold: Retraining AlphaFold2 Yields New Insights into Its Learning Mechanisms and Capacity for Generalization', *Nature Methods*, No. 21, 2024, pp. 1514–1524, https://doi.org/10.1038/s41592-024-02272-z.

Ahmed, N., Wahed, M. and Thompson, N.C., 'The Growing Influence of Industry in AI Research', *Science*, No. 379, 2023, pp. 884–886, https://doi.org/10.1126/science.ade2420.

Ahn, S., 'The Transformative Impact of Large Language Models on Medical Writing and Publishing: Current Applications, Challenges and Future Directions', *The Korean Journal of Physiology & Pharmacology : Official Journal of the Korean Physiological Society and the Korean Society of Pharmacology*, No. 28, The Korean Physiological Society and The Korean Society of Pharmacology, 2024, pp. 393–401, https://doi.org/10.4196/kjpp.2024.28.5.393.

Akdel, M., Pires, D.E.V., Pardo, E.P., Jänes, J., Zalevsky, A.O. et al., 'A Structural Biology Community Assessment of AlphaFold2 Applications', *Nature Structural & Molecular Biology*, No. 29, Nature Publishing Group, 2022, pp. 1056–1067, https://doi.org/10.1038/s41594-022-00849-w.

Albert, D. and Billinger, S., ‘Reproducing and Extending Experiments in Behavioral Strategy with Large Language Models’, *Academy of Management Proceedings*, No. 2025, Academy of Management, 2025, https://doi.org/10.5465/AMPROC.2025.299bp.

Alderson, T.R., Pritišanac, I., Kolarić, Đ., Moses, A.M. and Forman-Kay, J.D., ‘Systematic Identification of Conditionally Folded Intrinsically Disordered Regions by AlphaFold2’, *Proceedings of the National Academy of Sciences*, No. 120, 2023, https://doi.org/10.1073/pnas.2304302120.

Algaba, A., Mazijn, C., Holst, V., Tori, F., Wenmackers, S. et al., ‘Large Language Models Reflect Human Citation Patterns with a Heightened Citation Bias’, in: *Findings of the Association for Computational Linguistics: NAACL 2025*, Association for Computational Linguistics, Albuquerque, New Mexico, 2025, pp. 6829–6864, https://doi.org/10.18653/v1/2025.findings-naacl.381.

Allahyari, Z. and Oganov, A.R., ‘Coevolutionary Search for Optimal Materials in the Space of All Possible Compounds’, *npj Computational Materials*, No. 6, Nature Publishing Group, 2020, p. 55, https://doi.org/10.1038/s41524-020-0322-9.

ALLEA, ‘The European Code of Conduct for Research Integrity’, ALLEA - All European Academies, DE, 2023.

Alnaimat, F., AlSamhori, A.R.F., Hamdan, O., Seiil, B. and Qumar, A.B., ‘Perspectives of Artificial Intelligence Use for In-House Ethics Checks of Journal Submissions’, *Journal of Korean Medical Science*, No. 40, The Korean Academy of Medical Sciences, 2025, https://doi.org/10.3346/jkms.2025.40.e170.

Alvarado, R., ‘AI as an Epistemic Technology’, *Science and Engineering Ethics*, No. 29, 2023, p. 32, https://doi.org/10.1007/s11948-023-00451-3.

Alvarez, A., Caliskan, A., Crockett, M.J., Ho, S.S., Messeri, L. et al., ‘Science Communication with Generative AI’, *Nature Human Behaviour*, No. 8, Nature Publishing Group, 2024, pp. 625–627, https://doi.org/10.1038/s41562-024-01846-3.

Alverson, M., Baird, S.G., Murdock, R., Ho, (Enoch) Sin-Hang, Johnson, J. et al., ‘Generative Adversarial Networks and Diffusion Models in Material Discovery’, *Digital Discovery*, No. 3, 2024, pp. 62–80, https://doi.org/10.1039/d3dd00137g.

Andreeva, A., Howorth, D., Chothia, C., Kulesha, E. and Murzin, A.G., ‘SCOP2 Prototype: A New Approach to Protein Structure Mining’, *Nucleic Acids Research*, No. 42, 2014, pp. 310–314, https://doi.org/10.1093/nar/gkt1242.

Andreeva, A., Kulesha, E., Gough, J. and Murzin, A.G., ‘The SCOP Database in 2020: Expanded Classification of Representative Family and Superfamily Domains of Known Protein Structures’, *Nucleic Acids Research*, No. 48, 2020, pp. 376–382, https://doi.org/10.1093/nar/gkz1064.

Angelis, D., Sofos, F. and Karakasidis, T.E., ‘Artificial Intelligence in Physical Sciences: Symbolic Regression Trends and Perspectives’, *Archives of Computational Methods in Engineering*, No. 30, 2023, pp. 3845–3865, https://doi.org/10.1007/s11831-023-09922-z.

Anishchenko, I., Pellock, S.J., Chidyausiku, T.M., Ramelot, T.A., Ovchinnikov, S. et al., ‘De Novo Protein Design by Deep Network Hallucination’, *Nature*, No. 600, Nature Publishing Group, 2021, pp. 547–552, https://doi.org/10.1038/s41586-021-04184-w.

Antunes, L.M., Butler, K.T. and Grau-Crespo, R., ‘Crystal Structure Generation with Autoregressive Large Language Modeling’, *Nature Communications*, No. 15, 2024, https://doi.org/10.1038/s41467-024-54639-7.

Assael, Y., Sommerschield, T. and Prag, J., ‘Restoring Ancient Text Using Deep Learning: A Case Study on Greek Epigraphy’, arXiv, 2019, https://doi.org/10.48550/arXiv.1910.06262.

Assael, Y., Sommerschield, T., Shillingford, B., Bordbar, M., Pavlopoulos, J. et al., ‘Restoring and Attributing Ancient Texts Using Deep Neural Networks’, *Nature*, No. 603, 2022, pp. 280–283, https://doi.org/10.1038/s41586-022-04448-z.

ATLAS Collaboration, ‘Search for New Phenomena in Two-Body Invariant Mass Distributions Using Unsupervised Machine Learning for Anomaly Detection at √ s = 13 TeV with the ATLAS Detector’, *Physical Review Letters*, No. 132, American Physical Society, 2024.

Auer, S., ren, D’Souza, J., Farfar, K.E., Jaradeh, M.Y. et al., ‘Open Research Knowledge Graph: A Large-Scale Neuro-Symbolic Knowledge Organization System’, in: *Handbook on Neurosymbolic AI and Knowledge Graphs*, IOS Press, 2025, pp. 385–420, https://doi.org/10.3233/FAIA250216.

Auer, S., Stocker, M., Vogt, L., Fraumann, G. and Garatzogianni, A., ‘ORKG: Facilitating the Transfer of Research Results with the Open Research Knowledge Graph’, *Research Ideas and Outcomes*, No. 7, Pensoft Publishers, 2021, https://doi.org/10.3897/rio.7.e68513.

Augenstein, I., Baldwin, T., Cha, M., Chakraborty, T., Ciampaglia, G.L. et al., ‘Factuality Challenges in the Era of Large Language Models and Opportunities for Fact-Checking’, *Nature Machine Intelligence*, No. 6, Nature Publishing Group, 2024, pp. 852–863, https://doi.org/10.1038/s42256-024-00881-z.

Bacon, F., ‘Novum organum’, Clarendon Press, 1620.

Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S. et al., ‘Accurate Prediction of Protein Structures and Interactions Using a Three-Track Neural Network’, *Science*, No. 373, American Association for the Advancement of Science, 2021, pp. 871–876, https://doi.org/10.1126/science.abj8754.

Baek, M., McHugh, R., Anishchenko, I., Jiang, H., Baker, D. et al., ‘Accurate Prediction of Protein-Nucleic Acid Complexes Using RoseTTAFoldNA’, *Nature Methods*, No. 21, 2024, pp. 117–121, https://doi.org/10.1038/s41592-023-02086-5.

Bail, C.A., ‘Can Generative AI Improve Social Science?’, *Proceedings of the National Academy of Sciences*, No. 121, 2024, https://doi.org/10.1073/pnas.2314021121.

Baker, D. and Church, G., ‘Protein Design Meets Biosecurity’, *Science*, No. 383, American Association for the Advancement of Science, 2024, pp. 349–349, https://doi.org/10.1126/science.ado1671.

Bao, T., Zhao, Y., Mao, J. and Zhang, C., ‘Examining Linguistic Shifts in Academic Writing before and after the Launch of ChatGPT: A Study on Preprint Papers’, *Scientometrics*, 2025, https://doi.org/10.1007/s11192-025-05341-y.

Barman, K.G., Caron, S., Sullivan, E., de Regt, H.W., de Austri, R.R. et al., ‘Large Physics Models: Towards a Collaborative Approach with Large Language Models and Foundation Models’, arXiv, 2025, https://doi.org/10.48550/arXiv.2501.05382.

Barroso-Luque, L., Shuaibi, M., Fu, X., Wood, B.M., Dzamba, M. et al., ‘Open Materials 2024 (OMat24) Inorganic Materials Dataset and Models’, arXiv, 2024, https://doi.org/10.48550/arXiv.2410.12771.

Bartolomeis, P.D., Abad, J., Wang, G., Donhauser, K., Duch, R.M. et al., ‘Efficient Randomized Experiments Using Foundation Models’, arXiv, 2025, https://doi.org/10.48550/arXiv.2502.04262.

Batista, R.M. and Ross, J., ‘Words That Work: Using Language to Generate Hypotheses’, Social Science Research Network, Rochester, NY, 2024, https://doi.org/10.2139/ssrn.4926398.

Battleday, R.M. and Gershman, S.J., ‘Artificial Intelligence for Science: The Easy and Hard Problems’, arXiv, 2024, https://doi.org/10.48550/arXiv.2408.14508.

Beck, S., Poetz, M. and Sauermann, H., ‘How Will Artificial Intelligence (AI) Influence Openness and Collaboration in Science?’, Elephant in the Lab, 2022, https://doi.org/10.5281/zenodo.7211824.

Becker, B.A., Denny, P., Finnie-Ansley, J., Luxton-Reilly, A., Prather, J. et al., ‘Programming Is Hard - Or at Least It Used to Be: Educational Opportunities and Challenges of AI Code Generation’, in: *Proceedings of the 54th ACM Technical Symposium on Computer*

Science Education V. 1, SIGCSE 2023, Association for Computing Machinery, New York, NY, USA, 2023, pp. 500–506, https://doi.org/10.1145/3545945.3569759.

Bednarczyk, L., Reichenpfader, D., Gaudet-Blavignac, C., Ette, A.K., Zaghir, J. et al., ‘Scientific Evidence for Clinical Text Summarization Using Large Language Models: Scoping Review’, *Journal of Medical Internet Research*, No. 27, 2025, https://doi.org/10.2196/68998.

Beel, J., Kan, M.-Y. and Baumgart, M., ‘Evaluating Sakana’s AI Scientist for Autonomous Research: Wishful Thinking or an Emerging Reality Towards “Artificial Research Intelligence” (ARI)?’, arXiv, 2025, https://doi.org/10.48550/arXiv.2502.14297.

Bella, G., Helm, P., Koch, G. and Giunchiglia, F., ‘Tackling Language Modelling Bias in Support of Linguistic Diversity’, in: *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT ’24, Association for Computing Machinery, New York, NY, USA, 2024, pp. 562–572, https://doi.org/10.1145/3630106.3658925.

Bellat, M., Figueroa, J.D.O., Reeves, J.S., Taghizadeh-Mehrjardi, R., Tennie, C. et al., ‘Machine Learning Applications in Archaeological Practices: A Review’, arXiv, 2025, https://doi.org/10.48550/arXiv.2501.03840.

Bengio, Y., Mindermann, S., Privitera, D., Besiroglu, T., Bommasani, R. et al., ‘International AI Safety Report’, arXiv, 2025, https://doi.org/10.48550/arXiv.2501.17805.

Benhammou, Y., Etzion, E., Gadot, Y., Mizrachi, G., Lipschits, O. et al., ‘First Demonstration of Underground Muon Imaging at an Archaeological Site in Ancient Jerusalem’, https://doi.org/10.48550/arXiv.2504.03078.

Berens, P., Cranmer, K., Lawrence, N.D., Luxburg, U. von and Montgomery, J., ‘AI for Science: An Emerging Agenda’, arXiv, 2023, https://doi.org/10.48550/arXiv.2303.04217.

Berganzo-Besga, I., Orengo, H.A., Lumbreras, F., Carrero-Pazos, M., Fonte, J. et al., ‘Hybrid MSRM-Based Deep Learning and Multitemporal Sentinel 2-Based Machine Learning Algorithm Detects Near 10k Archaeological Tumuli in North-Western Iberia’, *Remote Sensing*, No. 13, 2021, p. 4181, https://doi.org/10.3390/rs13204181.

Bernardoni, S., Montanari, M. and Trojanis, R., ‘Open History Map’, *Archeologia e Calcolatori*, No. 28, 2017, pp. 539–548, https://doi.org/10.19282/AC.28.2.2017.44.

Bernasconi, E., Ceriani, M., Mecella, M. and Morvillo, A., ‘Automatic Knowledge Extraction from a Digital Library and Collaborative Validation’, in: Silvello, G., Corcho, O., Manghi, P., Di Nunzio, G.M., Golub, K. et al. (Eds.), *Linking Theory and Practice of Digital Libraries*, Springer International Publishing, Cham, 2022, pp. 480–484, https://doi.org/10.1007/978-3-031-16802-4_49.

Bertolini, L., Comte, V., Ruiz-Serra, V., Gierschmann, L., Orfei, L. et al., ‘Explainable AI in Cardiovascular Health’, Publications Office, Luxembourg, 2025.

Bhavsar, D., Duffy, L., Jo, H., Lokker, C., Haynes, R.B. et al., ‘Policies on Artificial Intelligence Chatbots among Academic Publishers: A Cross-Sectional Audit’, *Research Integrity and Peer Review*, No. 10, 2025, p. 1, https://doi.org/10.1186/s41073-025-00158-y.

Bi, Z., Xu, M., Tang, J. and Wang, X., ‘AI for Science in the Era of Large Language Models’, in: Li, J., and Liu, F. (Eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts*, Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 32–38, https://doi.org/10.18653/v1/2024.emnlp-tutorials.5.

Bianchini, S., Müller, M. and Pelletier, P., ‘Drivers and Barriers of AI Adoption and Use in Scientific Research’, arXiv, 2024, https://doi.org/10.48550/arXiv.2312.09843.

Binz, M., Alaniz, S., Roskies, A., Aczel, B., Bergstrom, C.T. et al., ‘How Should the Advancement of

Large Language Models Affect the Practice of Science?', *Proceedings of the National Academy of Sciences*, No. 122, 2025, https://doi.org/10.1073/pnas.2401227121.

Biron, L., 'Meet the Autonomous Lab of the Future', *Berkeley Lab News Center*, 2023, https://newscenter.lbl.gov/2023/04/17/meet-the-autonomous-lab-of-the-future.

Bloomfield, D., Pannu, J., Zhu, A.W., Ng, M.Y., Lewis, A. et al., 'AI and Biosecurity: The Need for Governance', *Science*, No. 385, American Association for the Advancement of Science, 2024, pp. 831–833, https://doi.org/10.1126/science.adq1977.

Bodel, J.P., 'Epigraphic Evidence: Ancient History From Inscriptions',Approaching the Ancient World, Taylor and Francis, Hoboken, 2012.

Boitreaud, J., Dent, J., McPartlon, M., Meier, J., Reis, V. et al., 'Chai-1: Decoding the Molecular Interactions of Life', bioRxiv, 2024, https://doi.org/10.1101/2024.10.10.615955.

Bolaños, F., Salatino, A., Osborne, F. and Motta, E., 'Artificial Intelligence for Literature Reviews: Opportunities and Challenges', *Artificial Intelligence Review*, No. 57, 2024, p. 259, https://doi.org/10.1007/s10462-024-10902-3.

Boutadjine, A., Harrag, F. and Shaalan, K., 'Human vs. Machine: A Comparative Study on the Detection of AI-Generated Content', *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, No. 24, 2025, pp. 1-26, https://doi.org/10.1145/3708889.

Broska, D. and McFarland, D., 'Engineering of Inquiry: The "Transformation" of Social Science through Generative AI', OSF, 2025, https://doi.org/10.31219/osf.io/uy4c7.

Bryda, G. and Sadowski, D., 'From Words to Themes: AI-Powered Qualitative Data Coding and Analysis', in: Ribeiro, J., Brandão, C., Ntsobi, M., Kasperiuniene, J., and Costa, A.P. (Eds.), *Computer Supported Qualitative Research*, Springer Nature Switzerland, Cham, 2024, pp. 309–345, https://doi.org/10.1007/978-3-031-65735-1_19.

Bühler, B., Bozkir, E., Goldberg, P., Sümer, Ö., D'Mello, S. et al., 'From the Lab to the Wild: Examining Generalizability of Video-Based Mind Wandering Detection', *International Journal of Artificial Intelligence in Education*, No. 35, 2025, pp. 823–857, https://doi.org/10.1007/s40593-024-00412-2.

Buriak, J.M., Akinwande, D., Artzi, N., Brinker, C.J., Burrows, C. et al., 'Best Practices for Using AI When Writing Scientific Manuscripts', *ACS Nano*, No. 17, American Chemical Society, 2023, pp. 4091–4093, https://doi.org/10.1021/acsnano.3c01544.

Burstein, J. and LaFlair, G.T., 'Where Assessment Validation and Responsible AI Meet', arXiv, 2024, https://doi.org/10.48550/arXiv.2411.02577.

Burton, J.W., Lopez-Lopez, E., Hechtlinger, S., Rahwan, Z., Aeschbach, S. et al., 'How Large Language Models Can Reshape Collective Intelligence', *Nature Human Behaviour*, No. 8, Nature Publishing Group, 2024, pp. 1643–1655, https://doi.org/10.1038/s41562-024-01959-9.

Callaway, E., 'Chemistry Nobel Goes to Developers of AlphaFold AI That Predicts Protein Structures', *Nature*, No. 634, 2024, pp. 525–526, https://doi.org/10.1038/d41586-024-03214-7.

Calza, E., Dalla Benetta, A., Kostić, U., Mitton, I., Moraschini, M. et al., 'Analytical insights into the global digital ecosystem (DGTES)', EUR 31538 EN, JRC132991, Publications Office of the European Union, Luxembourg, 2023.

Calza, E., Dalla Benetta, A., Kostić, U., Mitton, I., Vazquez-Prada Baillet, M. et al., 'A policy oriented analytical approach to map the digital ecosystem (DGTES)', EUR 31319 EN, JRC130799, Publications Office of the European Union, Luxembourg, 2022.

Canedo, D., Hipólito, J., Fonte, J., Dias, R., do Pereiro, T. et al., 'The Synergy between Artificial Intelligence, Remote Sensing, and Archaeological Fieldwork Validation', *Remote Sensing*, No. 16, Multidisciplinary Digital Publishing Institute, 2024, p. 1933, https://doi.org/10.3390/rs16111933.

Cano, F., 'Towards Responsible AI: Advances in Safety, Fairness, and Accountability of Autonomous Systems', arXiv, 2025, https://doi.org/10.48550/arXiv.2506.10192.

Cao, Z., Luo, X., Lv, J. and Wang, L., 'Space Group Informed Transformer for Crystalline Materials Generation', arXiv, 2024, https://doi.org/10.48550/arXiv.2403.15734.

Carpenter, J.M., Corvillón, A. and Shah, N.B., 'Enhancing Peer Review in Astronomy: A Machine Learning and Optimization Approach to Reviewer Assignments for ALMA', *Publications of the Astronomical Society of the Pacific*, No. 137, The Astronomical Society of the Pacific, 2025, https://doi.org/10.1088/1538-3873/adb5c1.

Carroll, S.R., Garba, I., Figueroa-Rodríguez, O.L., Holbrook, J., Lovett, R. et al., 'The CARE Principles for Indigenous Data Governance', *Data Science Journal*, No. 19, 2020, https://doi.org/10.5334/dsj-2020-043.

Carroll, S.R., Herczog, E., Hudson, M., Russell, K. and Stall, S., 'Operationalizing the CARE and FAIR Principles for Indigenous Data Futures', *Scientific Data*, No. 8, Nature Publishing Group, 2021, p. 108, https://doi.org/10.1038/s41597-021-00892-0.

Casini, L., Marchetti, N., Montanucci, A., Orrù, V. and Roccetti, M., 'A Human–AI Collaboration Workflow for Archaeological Sites Detection', *Scientific Reports*, No. 13, Nature Publishing Group, 2023, p. 8699, https://doi.org/10.1038/s41598-023-36015-5.

Chakravorti, T., Wang, X., Venkit, P.N., Koneru, S., Munger, K. et al., 'Social Scientists on the Role of AI in Research', arXiv, 2025, https://doi.org/10.48550/arXiv.2506.11255.

Chandonia, J.-M., Guan, L., Lin, S., Yu, C., Fox, N.K. et al., 'SCOPe: Improvements to the Structural Classification of Proteins - Extended Database to Facilitate Variant Interpretation and Machine Learning', *Nucleic Acids Research*, No. 50, 2022, pp. 553–559, https://doi.org/10.1093/nar/gkab1054.

Cheerkoot-Jalim, S. and Khedo, K.K., 'Literature-Based Discovery Approaches for Evidence-Based Healthcare: A Systematic Review', *Health and Technology*, No. 11, 2021, pp. 1205–1217, https://doi.org/10.1007/s12553-021-00605-y.

Cheetham, A.K. and Seshadri, R., 'Artificial Intelligence Driving Materials Discovery? Perspective on the Article: Scaling Deep Learning for Materials Discovery', *Chemistry of Materials*, No. 36, ACS Publications, 2024, pp. 3490–3495.

Chen, Q., Ho, Y.-J. (Ian), Sun, P. and Wang, D., 'The Philosopher's Stone for Science – The Catalyst Change of AI for Scientific Creativity', Social Science Research Network, Rochester, NY, 2024, https://doi.org/10.2139/ssrn.4749118.

Chen, Z., Chen, S., Ning, Y., Zhang, Q., Wang, B. et al., 'ScienceAgentBench: Toward Rigorous Assessment of Language Agents for Data-Driven Scientific Discovery', in: *The Thirteenth International Conference on Learning Representations*, 2024.

Chen, Z. and Zhang, J., 'Types of Patents and Driving Forces behind the Patent Growth in China', *Economic Modelling*, No. 80, 2019, pp. 294–302, https://doi.org/10.1016/j.econmod.2018.11.015.

Cheng, S., Zhao, X., Lu, G., Fang, J., Yu, Z. et al., 'FastFold: Reducing AlphaFold Training Time from 11 Days to 67 Hours', arXiv, 2023, https://doi.org/10.48550/arXiv.2203.00854.

Cheong, B.C., 'Transparency and Accountability in AI Systems: Safeguarding Wellbeing in the Age of Algorithmic Decision-Making', *Frontiers in Human Dynamics*, No. 6, Frontiers, 2024, https://doi.org/10.3389/fhumd.2024.1421273.

Chernozhukov, V., Hansen, C., Kallus, N., Spindler, M. and Syrgkanis, V., 'Applied Causal Inference Powered by ML and AI', arXiv, 2024, https://doi.org/10.48550/arXiv.2403.02467.

Chioma, V.N. and Lepe, O., 'The Intersection of Social Media and Artificial Intelligence: Examining Their Combined Influence on News Dissemination and Credibility in the Digital Age
', Social Science Research Network, Rochester, NY, 2024, https://doi.org/10.2139/ssrn.5036549.

Choudhary, K., Garrity, K.F., Reid, A.C.E., DeCost, B., Biacchi, A.J. et al., 'The Joint Automated Repository for Various Integrated Simulations (JARVIS) for Data-Driven Materials Design', *npj Computational Materials*, No. 6, Nature Publishing Group, 2020, p. 173, https://doi.org/10.1038/s41524-020-00440-1.

Chowdhury, R., Bouatta, N., Biswas, S., Floristean, C., Kharkar, A. et al., 'Single-Sequence Protein Structure Prediction Using a Language Model and Deep Learning', *Nature Biotechnology*, No. 40, 2022, pp. 1617–1623, https://doi.org/10.1038/s41587-022-01432-w.

Chowdhury, S., Joel-Edgar, S., Dey, P.K., Bhattacharya, S. and Kharlamov, A., 'Embedding Transparency in Artificial Intelligence Machine Learning Models: Managerial Implications on Predicting and Explaining Employee Turnover', *The International Journal of Human Resource Management*, No. 34, Routledge, 2023, pp. 2732–2764, https://doi.org/10.1080/09585192.2022.2066981.

Christou, P.A., 'The Use of Artificial Intelligence (AI) in Qualitative Research for Theory Development', *The Qualitative Report*, No. 28, 2023, pp. 2739–2755.

Coeckelbergh, M., 'AI and Epistemic Agency: How AI Influences Belief Revision and Its Normative Implications', *Social Epistemology*, Informa UK Limited, 2025, pp. 1–13, https://doi.org/10.1080/02691728.2025.2466164.

Cohen, I. and Peled, D., 'LLM-Based Scheme for Synthesis of Formal Verification Algorithms', in: Steffen, B. (Ed.), *Bridging the Gap Between AI and Reality*, Springer Nature Switzerland, Cham, 2025, pp. 167–182, https://doi.org/10.1007/978-3-031-75434-0_11.

Conte, R., Gilbert, N., Bonelli, G., Cioffi-Revilla, C., Deffuant, G. et al., 'Manifesto of Computational Social Science', *The European Physical Journal Special Topics*, No. 214, 2012, pp. 325–346, https://doi.org/10.1140/epjst/e2012-01697-8.

Conti, S., 'Artificial Intelligence for Weather Forecasting', *Nature Reviews Electrical Engineering*, No. 1, Nature Publishing Group, 2024, https://doi.org/10.1038/s44287-023-00009-2.

Cornelio, C., Dash, S., Austel, V., Josephson, T.R., Goncalves, J. et al., 'Combining Data and Theory for Derivable Scientific Discovery with AI-Descartes', *Nature Communications*, No. 14, Nature Publishing Group, 2023, p. 1777, https://doi.org/10.1038/s41467-023-37236-y.

Court, C.J., Yildirim, B., Jain, A. and Cole, J.M., '3-D Inorganic Crystal Structure Generation and Property Prediction via Representation Learning', *Journal of Chemical Information and Modeling*, No. 60, American Chemical Society, 2020, pp. 4518–4535, https://doi.org/10.1021/acs.jcim.0c00464.

Curtarolo, S., Hart, G.L.W., Nardelli, M.B., Mingo, N., Sanvito, S. et al., 'The High-Throughput Highway to Computational Materials Design', *Nature Materials*, No. 12, Nature Publishing Group, 2013, pp. 191–201, https://doi.org/10.1038/nmat3568.

Curtarolo, S., Setyawan, W., Hart, G.L.W., Jahnatek, M., Chepulskii, R.V. et al., 'AFLOW: An Automatic Framework for High-Throughput Materials Discovery', *Computational Materials Science*, No. 58, 2012, pp. 218–226, https://doi.org/10.1016/j.commatsci.2012.02.005.

Czako, Z., Sebestyen, G. and Hangan, A., 'AutomaticAI – A Hybrid Approach for Automatic Artificial Intelligence Algorithm Selection and Hyperparameter Tuning', *Expert Systems with Applications*, No. 182, 2021, https://doi.org/10.1016/j.eswa.2021.115225.

Dagdelen, J., Dunn, A., Lee, S., Walker, N., Rosen, A.S. et al., 'Structured Information Extraction from Scientific Text with Large Language Models', *Nature Communications*, No. 15, Nature Publishing Group, 2024, p. 1418, https://doi.org/10.1038/s41467-024-45563-x.

Dang, X., Li, Y., Papadakis, M., Klein, J., Bissyandé, T.F. et al., 'Test Input Prioritization for Machine Learning Classifiers', *IEEE Transactions on Software Engineering*, No. 50, 2024, pp. 413–442, https://doi.org/10.1109/TSE.2024.3350019.

Dara, S., Dhamercherla, S., Jadav, S.S., Babu, C.M. and Ahsan, M.J., 'Machine Learning in Drug Discovery: A Review', *Artificial Intelligence Review*, No. 55, 2022, pp. 1947–1999, https://doi.org/10.1007/s10462-021-10058-4.

Davies, A., Veličković, P., Buesing, L., Blackwell, S., Zheng, D. et al., 'Advancing Mathematics by Guiding Human Intuition with AI', *Nature*, No. 600, Nature Publishing Group, 2021, pp. 70–74, https://doi.org/10.1038/s41586-021-04086-x.

Davies, D.W., Butler, K.T., Jackson, A.J., Morris, A., Frost, J.M. et al., 'Computational Screening of All Stoichiometric Inorganic Materials', *Chem*, No. 1, Elsevier, 2016, pp. 617–627, https://doi.org/10.1016/j.chempr.2016.09.010.

De La Vega, F.M., Chowdhury, S., Moore, B., Frise, E., McCarthy, J. et al., 'Artificial Intelligence Enables Comprehensive Genome Interpretation and Nomination of Candidate Diagnoses for Rare Genetic Diseases', *Genome Medicine*, No. 13, 2021, p. 153, https://doi.org/10.1186/s13073-021-00965-0.

DeGroat, W., Mendhe, D., Bhusari, A., Abdelhalim, H., Zeeshan, S. et al., 'IntelliGenes: A Novel Machine Learning Pipeline for Biomarker Discovery and Predictive Analysis Using Multi-Genomic Profiles', *Bioinformatics*, No. 39, 2023, https://doi.org/10.1093/bioinformatics/btad755.

Delgado-Licona, F. and Abolhasani, M., 'Research Acceleration in Self-Driving Labs: Technological Roadmap toward Accelerated Materials and Molecular Discovery', *Advanced Intelligent Systems*, No. 5, 2023, https://doi.org/10.1002/aisy.202200331.

Desai, S., Addamane, S., Tsao, J.Y., Brener, I., Swiler, L.P. et al., 'AutoSciLab: A Self-Driving Laboratory for Interpretable Scientific Discovery', *Proceedings of the AAAI Conference on Artificial Intelligence*, No. 39, 2025, pp. 146–154, https://doi.org/10.1609/aaai.v39i1.31990.

Dewey, J., 'How we think', D.C. Heath & Co., 1910.

Dibia, V., 'LIDA: A Tool for Automatic Generation of Grammar-Agnostic Visualizations and Infographics Using Large Language Models', in: Bollegala, D., Huang, R., and Ritter, A. (Eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Association for Computational Linguistics, Toronto, Canada, 2023, pp. 113–126, https://doi.org/10.18653/v1/2023.acl-demo.11.

Dierickx, L., van Dalen, A., Opdahl, A.L. and Lindén, C.-G., 'Striking the Balance in Using LLMs for Fact-Checking: A Narrative Literature Review', in: Preuss, M., Leszkiewicz, A., Boucher, J.-C., Fridman, O., and Stampe, L. (Eds.), *Disinformation in Open Online Media*, Springer Nature Switzerland, Cham, 2024, pp. 1–15, https://doi.org/10.1007/978-3-031-71210-4_1.

Ding, R., Han, S. and Zhang, D., 'InsightPilot: An LLM-Empowered Automated Data Exploration System', in: *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

Ding, X. and Li, J., 'Incentives for Innovation in China: Building an Innovative Economy', Routledge, London, 2015, https://doi.org/10.4324/9781315726717.

Dixit, V., Hussain, N., Basak, S., Atturu, D., Mitra, D. et al., 'Deep Learning in Archiving Indus Script and Motif Information', *Journal of Computer Applications in Archaeology*, No. 8, 2025, https://doi.org/10.5334/jcaa.175.

Dobson, C.M., 'Protein Folding and Misfolding', *Nature*, No. 426, Nature Publishing Group, 2003, pp. 884–890, https://doi.org/10.1038/nature02261.

Dobson, L., Szekeres, L.I., Gerdán, C., Langó, T., Zeke, A. et al., 'TmAlphaFold Database: Membrane Localization and Evaluation of AlphaFold2

Predicted Alpha-Helical Transmembrane Protein Structures', *Nucleic Acids Research*, No. 51, 2023, pp. 517–522, https://doi.org/10.1093/nar/gkac928.

Doshi, A.R. and Hauser, O.P., 'Generative Artificial Intelligence Enhances Creativity but Reduces the Diversity of Novel Content', *Science Advances*, No. 10, 2024, https://doi.org/10.1126/sciadv.adn5290.

Dou, S., Jia, H., Wu, S., Zheng, H., Zhou, W. et al., 'What's Wrong with Your Code Generated by Large Language Models? An Extensive Study', arXiv, 2024, https://doi.org/10.48550/arXiv.2407.06153.

Drosos, I., Sarkar, A., Xu, X., Negreanu, C., Rintel, S. et al., '"It's like a Rubber Duck That Talks Back": Understanding Generative AI-Assisted Data Analysis Workflows through a Participatory Prompting Study', in: *Proceedings of the 3rd Annual Meeting of the Symposium on Human-Computer Interaction for Work*, ACM, Newcastle upon Tyne United Kingdom, 2024, pp. 1–21, https://doi.org/10.1145/3663384.3663389.

Duede, E., 'Deep Learning Opacity in Scientific Discovery', *Philosophy of Science*, No. 90, 2023, pp. 1089–1099, https://doi.org/10.1017/psa.2023.8.

Dworschak, F., Dietze, S., Wittmann, M., Schleich, B. and Wartzack, S., 'Reinforcement Learning for Engineering Design Automation', *Advanced Engineering Informatics*, No. 52, 2022, https://doi.org/10.1016/j.aei.2022.101612.

Edich, M., Briggs, D.C., Kippes, O., Gao, Y. and Thorn, A., 'The Impact of AlphaFold2 on Experimental Structure Solution', *Faraday Discussions*, No. 240, The Royal Society of Chemistry, 2022, pp. 184–195, https://doi.org/10.1039/D2FD00072E.

Eger, S., Cao, Y., D'Souza, J., Geiger, A., Greisinger, C. et al., 'Transforming Science with Large Language Models: A Survey on AI-Assisted Scientific Discovery, Experimentation, Content Generation, and Evaluation', arXiv, 2025, https://doi.org/10.48550/arXiv.2502.05151.

Elali, F.R. and Rachid, L.N., 'AI-Generated Research Paper Fabrication and Plagiarism in the Scientific Community', *Patterns*, No. 4, Elsevier, 2023, https://doi.org/10.1016/j.patter.2023.100706.

Elkhatat, A.M., Elsaid, K. and Almeer, S., 'Evaluating the Efficacy of AI Content Detection Tools in Differentiating between Human and AI-Generated Text', *International Journal for Educational Integrity*, No. 19, 2023, p. 17, https://doi.org/10.1007/s40979-023-00140-5.

Emmerich, S., Baumgart, U., Beckmann, S., Steidel, S. and Burger, M., 'AI-Based Surrogate Modeling for Highly Efficient Soil-Tool Simulation', in: Berns, K., Dreßler, K., Kalmar, R., Stephan, N., Teutsch, R. et al. (Eds.), *Commercial Vehicle Technology 2024*, Springer Fachmedien, Wiesbaden, 2024, pp. 205–221, https://doi.org/10.1007/978-3-658-45699-3_12.

Eriksson, M., Purificato, E., Noroozian, A., Vinagre, J., Chaslot, G. et al., 'Can We Trust AI Benchmarks? An Interdisciplinary Review of Current Issues in AI Evaluation', arXiv, 2025, https://doi.org/10.48550/arXiv.2502.06559.

European Commission: Directorate General for Research and Innovation, 'Living Guidelines on the Responsible Use of Generative AI in Research (Second Version)' (ERA Forum Stakeholders's document), Publications Office, Luxembourg, 2025.

European Commission: Directorate General for Research and Innovation, 'Horizon Europe Strategic Plan 2025-2027', Publications Office, Luxembourg, 2024.

European Commission: Directorate General for Research and Innovation, 'AI in science: harnessing the power of AI to accelerate discovery and foster innovation : policy brief.', Publications Office of the European Union, Luxembourg, 2023.

European Commission: Directorate General for Research and Innovation, Arranz, D., Bianchini, S., Di Girolamo, V. and Ravet, J., 'Trends in the Use of AI in Science - A Bibliometric Analysis', Publications Office, Luxembourg, 2023.

European Commission: Directorate General for Research and Innovation, Group of Chief Scientific Advisors and King, R., 'Infrastructures for AI in Science – Mutual Learning Exercise on National Policies for AI in Science – Second Thematic Report', Publications Office, Luxembourg, 2025.

European Commission: Directorate General for Research and Innovation and Groznik, V., 'Talent for AI in Science – Mutual Learning Exercise on National Policies on AI in Science – Third Thematic Report', Publications Office, Luxembourg, 2025.

European Commission: Directorate General for Research and Innovation and Montgomery, J., 'Framework Conditions and Funding for AI in Science – Mutual Learning Exercise on National Policies for AI in Science – First Thematic Report', Publications Office, Luxembourg, 2025.

European Commission: Directorate-General for Communications Networks, Content and Technology and High-Level Expert Group on Artificial Intelligence, 'Ethics Guidelines for Trustworthy AI', Publications Office, Luxembourg, 2019.

Evans, R., O'Neill, M., Pritzel, A., Antropova, N., Senior, A. et al., 'Protein Complex Prediction with AlphaFold-Multimer', bioRxiv, 2022, https://doi.org/10.1101/2021.10.04.463034.

Feng, G.C., 'Best Practices for Responsibly Using AI Tools in Social Sciences Research', *Cogent Social Sciences*, No. 10, Cogent OA, 2024, https://doi.org/10.1080/23311886.2024.2420484.

Flam-Shepherd, D. and Aspuru-Guzik, A., 'Language Models Can Generate Molecules, Materials, and Protein Binding Sites Directly in Three Dimensions as XYZ, CIF, and PDB Files', arXiv, 2023, https://doi.org/10.48550/arXiv.2305.05708.

Fluke, C.J. and Jacobs, C., 'Surveying the Reach and Maturity of Machine Learning and Artificial Intelligence in Astronomy', *WIREs Data Mining and Knowledge Discovery*, No. 10, 2020, p. e1349, https://doi.org/10.1002/widm.1349.

Forer, L. and Hope, T., 'Inferring Scientific Cross-Document Coreference and Hierarchy with Definition-Augmented Relational Reasoning', arXiv, 2024, https://doi.org/10.48550/arXiv.2409.15113.

Fortson, L., Crowston, K., Kloetzer, L. and Ponti, M., 'Artificial Intelligence and the Future of Citizen Science', *Citizen Science: Theory and Practice*, No. 9, 2024, https://doi.org/10.5334/cstp.812.

Fox, N.K., Brenner, S.E. and Chandonia, J.-M., 'SCOPe: Structural Classification of Proteins—Extended, Integrating SCOP and ASTRAL Data and Classification of New Structures', *Nucleic Acids Research*, No. 42, 2014, pp. 304–309, https://doi.org/10.1093/nar/gkt1240.

Freeze, J.G., Kelly, H.R. and Batista, V.S., 'Search for Catalysts by Inverse Design: Artificial Intelligence, Mountain Climbers, and Alchemists', *Chemical Reviews*, No. 119, American Chemical Society, 2019, pp. 6595–6612, https://doi.org/10.1021/acs.chemrev.8b00759.

Fudan University and Shanghai Academy of AI for Science, 'AI for Science',2025.

Gabriel, I., 'Artificial Intelligence, Values, and Alignment', *Minds and Machines*, No. 30, 2020, pp. 411–437, https://doi.org/10.1007/s11023-020-09539-2.

Galli, C., Moretti, C. and Calciolari, E., 'Intelligent Summaries: Will Artificial Intelligence Mark the Finale for Biomedical Literature Reviews?', *Learned Publishing*, No. 38, 2024, https://doi.org/10.1002/leap.1648.

Gandini, A., Keeling, S. and Reviglio, U., 'Conceptualising the "Algorithmic Public Opinion": Public Opinion Formation in the Digital Age', *Dialogues on Digital Society*, SAGE Publications, 2025, https://doi.org/10.1177/29768640251323147.

Gao, S., Fang, A., Huang, Y., Giunchiglia, V., Noori, A. et al., 'Empowering Biomedical Discovery with AI Agents', *Cell*, No. 187, Elsevier, 2024, pp. 6125–6151, https://doi.org/10.1016/j.cell.2024.09.022.

Gattiglia, G., ‘Managing Artificial Intelligence in Archeology. An Overview’, *Journal of Cultural Heritage*, No. 71, 2025, pp. 225–233, https://doi.org/10.1016/j.culher.2024.11.020.

Gebauer, N., Gastegger, M. and Schütt, K., ‘Symmetry-Adapted Generation of 3d Point Sets for the Targeted Discovery of Molecules’, in: *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2019.

Geng, M. and Trotta, R., ‘Is ChatGPT Transforming Academics’ Writing Style?’, arXiv, 2024, https://doi.org/10.48550/arXiv.2404.08627.

Ghafarollahi, A. and Buehler, M.J., ‘SciAgents: Automating Scientific Discovery Through Bioinspired Multi-Agent Intelligent Graph Reasoning’, *Advanced Materials*, 2024, https://doi.org/10.1002/adma.202413523.

Gharaibeh, M., Alzu’bi, D., Abdullah, M., Hmeidi, I., Al Nasar, M.R. et al., ‘Radiology Imaging Scans for Early Diagnosis of Kidney Tumors: A Review of Data Analytics-Based Machine Learning and Deep Learning Approaches’, *Big Data and Cognitive Computing*, No. 6, Multidisciplinary Digital Publishing Institute, 2022, p. 29, https://doi.org/10.3390/bdcc6010029.

Giarelis, N., Mastrokostas, C. and Karacapilidis, N., ‘A Unified LLM-KG Framework to Assist Fact-Checking in Public Deliberation’, in: *Proceedings of the First Workshop on Language-Driven Deliberation Technology (DELITE) @ LREC-COLING 2024*, ELRA and ICCL, Torino, Italia, 2024, pp. 13–19.

Giere, R.N., ‘Understanding scientific reasoning’, Holt, Rinehart and Winston, 1979.

Gil, Y., Garijo, D., Khider, D., Knoblock, C.A., Ratnakar, V. et al., ‘Artificial Intelligence for Modeling Complex Systems: Taming the Complexity of Expert Models to Improve Decision Making’, *ACM Trans. Interact. Intell. Syst.*, No. 11, 2021, pp. 1-49, https://doi.org/10.1145/3453172.

Glickman, M. and Zhang, Y., ‘AI and Generative AI for Research Discovery and Summarization’, arXiv, 2024, https://doi.org/10.48550/arXiv.2401.06795.

Gomez, E., Porcaro, L., Frau Amar, P. and Vinagre, J., ‘Diversity in Artificial Intelligence Conferences’, EUR 31968 EN, JRC137550, Publications Office of the European Union, Luxembourg, 2024.

Goretti, L., Küsters ,Anselm, Morgana ,Stella and and Sharma, A., ‘Challenges and Opportunities of Using Generative AI for Research: Opening the Discussion’, *The International Spectator*, No. 60, Routledge, 2025, pp. 1–16, https://doi.org/10.1080/03932729.2024.2435176.

Gottweis, J., Weng, W.-H., Daryin, A., Tu, T., Palepu, A. et al., ‘Towards an AI Co-Scientist’, arXiv, 2025, https://doi.org/10.48550/arXiv.2502.18864.

Goyal, M. and Mahmoud, Q.H., ‘A Systematic Review of Synthetic Data Generation Techniques Using Generative AI’, *Electronics*, No. 13, Multidisciplinary Digital Publishing Institute, 2024, p. 3509, https://doi.org/10.3390/electronics13173509.

Gray, A., ‘ChatGPT “Contamination”: Estimating the Prevalence of LLMs in the Scholarly Literature’, arXiv, 2024, https://doi.org/10.48550/arXiv.2403.16887.

Greenhill, S., Rana, S., Gupta, S., Vellanki, P. and Venkatesh, S., ‘Bayesian Optimization for Adaptive Experimental Design: A Review’, *IEEE Access*, No. 8, 2020, pp. 13937–13948, https://doi.org/10.1109/ACCESS.2020.2966228.

Gruver, N., Sriram, A., Madotto, A., Wilson, A.G., Zitnick, C.L. et al., ‘Fine-Tuned Language Models Generate Stable Inorganic Materials as Text’, in: *The Twelfth International Conference on Learning Representations*, 2023.

Gu, T., Wang, J., Zhang, Z. and Li, H., ‘LLMs Can Realize Combinatorial Creativity: Generating Creative Ideas via LLMs for Scientific Research’, arXiv, 2025, https://doi.org/10.48550/arXiv.2412.14141.

Gu, X. and Krenn, M., ‘Forecasting High-Impact Research Topics via Machine Learning on Evolving Knowledge Graphs’, *Machine Learning: Science and Technology*, No. 6, IOP Publishing, 2025a, https://doi.org/10.1088/2632-2153/add6ef.

Gu, X. and Krenn, M., ‘Interesting Scientific Idea Generation Using Knowledge Graphs and LLMs: Evaluations with 100 Research Group Leaders’, arXiv, 2025b, https://doi.org/10.48550/arXiv.2405.17044.

Gupta, T. and Pruthi, D., ‘All That Glitters Is Not Novel: Plagiarism in AI Generated Research’, arXiv, 2025, https://doi.org/10.48550/arXiv.2502.16487.

Gürcan, Ö., ‘LLM-Augmented Agent-Based Modelling for Social Simulations: Challenges and Opportunities’, in: *Proceedings of the Third International Conference on Hybrid Human-Artificial Intelligence*, IOS Press, 2024.

Guyot, A., Hubert-Moy, L. and Lorho, T., ‘Detecting Neolithic Burial Mounds from LiDAR-Derived Elevation Data Using a Multi-Scale Approach and Machine Learning Techniques’, *Remote Sensing*, No. 10, Multidisciplinary Digital Publishing Institute, 2018, p. 225, https://doi.org/10.3390/rs10020225.

Hanson, B., Stall, S., Cutcher-Gershenfeld, J., Vrouwenvelder, K., Wirz, C. et al., ‘Garbage in, Garbage out: Mitigating Risks and Maximizing Benefits of AI in Research’, *Nature*, No. 623, 2023, pp. 28–31, https://doi.org/10.1038/d41586-023-03316-8.

Hao, Q., Xu, F., Li, Y. and Evans, J., ‘AI Expands Scientists’ Impact but Contracts Science’s Focus’, arXiv, 2024, https://doi.org/10.48550/ARXIV.2412.07727.

He, J., Feng, W., Min, Y., Yi, J., Tang, K. et al., ‘Control Risk for Potential Misuse of Artificial Intelligence in Science’, arXiv, 2023, https://doi.org/10.48550/arXiv.2312.06632.

He, K., Zhang, X., Ren, S. and Sun, J., ‘Deep Residual Learning for Image Recognition’, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

Heinzinger, M., Weissenow, K., Sanchez, J.G., Henkel, A., Mirdita, M. et al., ‘Bilingual Language Model for Protein Sequence and Structure’, *NAR Genomics and Bioinformatics*, No. 6, 2024, https://doi.org/10.1093/nargab/lqae150.

Hempel, C.G., ‘Philosophy of natural science’, Prentice-Hall, 1966.

Henry, S. and McInnes, B.T., ‘Literature Based Discovery: Models, Methods, and Trends’, *Journal of Biomedical Informatics*, No. 74, 2017, pp. 20–32, https://doi.org/10.1016/j.jbi.2017.08.011.

Himanen, L., Geurts, A., Foster, A.S. and Rinke, P., ‘Data-Driven Materials Science: Status, Challenges, and Perspectives’, *Advanced Science*, No. 6, 2019, https://doi.org/10.1002/advs.201900808.

Hoffmann, J., Maestrati, L., Sawada, Y., Tang, J., Sellier, J.M. et al., ‘Data-Driven Approach to Encoding and Decoding 3-D Crystal Structures’, arXiv, 2019, https://doi.org/10.48550/arXiv.1909.00949.

Hong, Y., Lian, J., Xu, L., Min, J., Wang, Y. et al., ‘Statistical Perspectives on Reliability of Artificial Intelligence Systems’, *Quality Engineering*, No. 35, Taylor & Francis, 2023, pp. 56–78, https://doi.org/10.1080/08982112.2022.2089854.

Hosseini, M. and Horbach, S.P.J.M., ‘Fighting Reviewer Fatigue or Amplifying Bias? Considerations and Recommendations for Use of ChatGPT and Other Large Language Models in Scholarly Peer Review’, *Research Integrity and Peer Review*, No. 8, 2023, p. 4, https://doi.org/10.1186/s41073-023-00133-5.

Hou, Y. and Huang, J., ‘Natural Language Processing for Social Science Research: A Comprehensive Review’, *Chinese Journal of Sociology*, No. 11, SAGE Publications, 2025, pp. 121–157, https://doi.org/10.1177/2057150X241306780.

Hsu, C.-C., Bransom, E., Sparks, J., Kuehl, B., Tan, C. et al., ‘CHIME: LLM-Assisted Hierarchical Organization of Scientific Studies for Literature Review Support’, in: *Findings of the Association for Computational Linguistics: ACL 2024*, Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 118–132, https://doi.org/10.18653/v1/2024.findings-acl.8.

Hu, B., Xia, J., Zheng, J., Tan, C., Huang, Y. et al., 'Protein Language Models and Structure Prediction: Connection and Progression', arXiv, 2022, https://doi.org/10.48550/arXiv.2211.16742.

Hu, T., Kyrychenko, Y., Rathje, S., Collier, N., Van Der Linden, S. et al., 'Generative Language Models Exhibit Social Identity Biases', *Nature Computational Science*, No. 5, 2024, pp. 65–75, https://doi.org/10.1038/s43588-024-00741-1.

Hu, X., Gao, H., Agafari, T., Zhang, M.Q. and Cong, R., 'How and When Artificial Intelligence Adoption Promotes Employee Knowledge Sharing? The Role of Paradoxical Leadership and Technophilia', *Frontiers in Psychology*, No. 16, Frontiers, 2025, https://doi.org/10.3389/fpsyg.2025.1573587.

Hu, X., Zhao, Z., Wei, S., Chai, Z., Ma, Q. et al., 'InfiAgent-DABench: Evaluating Agents on Data Analysis Tasks', in: *Proceedings of the 41st International Conference on Machine Learning*, PMLR, 2024, pp. 19544–19572.

Hu, Y., Liu, D., Wang, Q., Yu, C., Xu, C. et al., 'Automating Intervention Discovery from Scientific Literature: A Progressive Ontology Prompting and Dual-LLM Framework', arXiv, 2025, https://doi.org/10.48550/arXiv.2409.00054.

Huang, F., Kwak, H. and An, J., 'Is ChatGPT Better than Human Annotators? Potential and Limitations of ChatGPT in Explaining Implicit Hate Speech', in: *Companion Proceedings of the ACM Web Conference 2023*, 2023, pp. 294–297, https://doi.org/10.1145/3543873.3587368.

Huang, K., Hu, T., Cai, J., Pan, X., Hou, Y. et al., 'Artificial Intelligence in Astronomical Optical Telescopes: Present Status and Future Perspectives', *Universe*, No. 10, Multidisciplinary Digital Publishing Institute, 2024, p. 210, https://doi.org/10.3390/universe10050210.

Huang, K., Jin, Y., Li, R., Li, M.Y., Candès, E. et al., 'Automated Hypothesis Validation with Agentic Sequential Falsifications', arXiv, 2025, https://doi.org/10.48550/arXiv.2502.09858.

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z. et al., 'A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions', *ACM Trans. Inf. Syst.*, No. 43, 2025, pp. 1-55, https://doi.org/10.1145/3703155.

Huang, Q., Vora, J., Liang, P. and Leskovec, J., 'MLAgentBench: Evaluating Language Agents on Machine Learning Experimentation', in: *Proceedings of the 41st International Conference on Machine Learning*, PMLR, 2024, pp. 20271–20309.

Huang, Y., Chen, Y., Zhang, H., Li, K., Fang, M. et al., 'Deep Research Agents: A Systematic Examination And Roadmap', arXiv, 2025, https://doi.org/10.48550/arXiv.2506.18096.

Huerta, E.A., Blaiszik, B., Brinson, L.C., Bouchard, K.E., Diaz, D. et al., 'FAIR for AI: An Interdisciplinary and International Community Building Perspective', *Scientific Data*, No. 10, Nature Publishing Group, 2023, p. 487, https://doi.org/10.1038/s41597-023-02298-6.

Hunklinger, A. and Ferruz, N., 'Toward the Explainability of Protein Language Models for Sequence Design', arXiv, 2025, https://doi.org/10.48550/arXiv.2506.19532.

Ifargan, T., Hafner, L., Kern, M., Alcalay, O. and Kishony, R., 'Autonomous LLM-Driven Research — from Data to Human-Verifiable Research Papers', *NEJM AI*, No. 2, Massachusetts Medical Society, 2025, https://doi.org/10.1056/AIoa2400555.

Ison, J., Ienasescu, H., Chmura, P., Rydza, E., Ménager, H. et al., 'The Bio.Tools Registry of Software Tools and Data Resources for the Life Sciences', *Genome Biology*, No. 20, 2019, https://doi.org/10.1186/s13059-019-1772-6.

Jaakkola, M., 'Academic AI Literacy: Artificial Intelligence in Scholarly Writing, Editing, and Publishing', NordMedia Network, 2024.

Jacobs, R., Mayeshiba, T., Afflerbach, B., Miles, L., Williams, M. et al., 'The Materials Simulation Toolkit for Machine Learning (MAST-ML): An Automated Open Source Toolkit to Accelerate Data-Driven Materials Research', *Computational Materials Science*, No. 176, 2020, https://doi.org/10.1016/j.commatsci.2020.109544.

Jacquot, R., Sève, P., Jackson, T.L., Wang, T., Duclos, A. et al., 'Diagnosis, Classification, and Assessment of the Underlying Etiology of Uveitis by Artificial Intelligence: A Systematic Review', *Journal of Clinical Medicine*, No. 12, Multidisciplinary Digital Publishing Institute, 2023, p. 3746, https://doi.org/10.3390/jcm12113746.

Jain, A., Ong, S.P., Hautier, G., Chen, W., Richards, W.D. et al., 'Commentary: The Materials Project: A Materials Genome Approach to Accelerating Materials Innovation', *APL Materials*, No. 1, 2013, https://doi.org/10.1063/1.4812323.

Jalaian, B. and Bastian, N.D., 'Neurosymbolic AI in Cybersecurity: Bridging Pattern Recognition and Symbolic Reasoning', in: *MILCOM 2023 - 2023 IEEE Military Communications Conference (MILCOM)*, 2023, pp. 268–273, https://doi.org/10.1109/MILCOM58377.2023.10356283.

Jejjala, V., Kar, A. and Parrikar, O., 'Deep Learning the Hyperbolic Volume of a Knot', *Physics Letters B*, No. 799, 2019, https://doi.org/10.1016/j.physletb.2019.135033.

Jeon, J., Kim, L. and Park, J., 'The Ethics of Generative AI in Social Science Research: A Qualitative Approach for Institutionally Grounded AI Research Ethics', *Technology in Society*, No. 81, 2025, https://doi.org/10.1016/j.techsoc.2025.102836.

Jiang, J., Huang, Z.-G., Grebogi, C. and Lai, Y.-C., 'Predicting Extreme Events from Data Using Deep Machine Learning: When and Where', *Physical Review Research*, No. 4, American Physical Society, 2022, https://doi.org/10.1103/PhysRevResearch.4.023028.

Jiang, X., Xue, D., Bai, Y., Wang, W.Y., Liu, J. et al., 'AI4Materials: Transforming the Landscape of Materials Science and Enigneering', *Review of Materials Research*, No. 1, 2025, p. 100010, https://doi.org/10.1016/j.revmat.2025.100010.

Jiao, L., Song, X., You, C., Liu, X., Li, L. et al., 'AI Meets Physics: A Comprehensive Survey', *Artificial Intelligence Review*, No. 57, 2024, p. 256, https://doi.org/10.1007/s10462-024-10874-4.

Jiao, R., Huang, W., Lin, P., Han, J., Chen, P. et al., 'Crystal Structure Prediction by Joint Equivariant Diffusion', *Advances in Neural Information Processing Systems*, No. 36, 2023, pp. 17464–17497.

Jones, A., Kuehnert, J., Fraccaro, P., Meuriot, O., Ishikawa, T. et al., 'AI for Climate Impacts: Applications in Flood Risk', *npj Climate and Atmospheric Science*, No. 6, Nature Publishing Group, 2023, p. 63, https://doi.org/10.1038/s41612-023-00388-1.

Jorge, C., 'The Importance of Archaeology in Preserving Cultural Heritage and History', *Anthropology*, No. 12, Longdom Publishing S.L, 2024, pp. 1–2, https://doi.org/10.35248/2332-0915.24.12.242.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M. et al., 'Highly Accurate Protein Structure Prediction with AlphaFold', *Nature*, No. 596, Nature Publishing Group, 2021, pp. 583–589, https://doi.org/10.1038/s41586-021-03819-2.

Jurowetzki, R., Hain, D.S., Wirtz, K. and Bianchini, S., 'The Private Sector Is Hoarding AI Researchers: What Implications for Science?', *AI & SOCIETY*, No. 40, 2025, pp. 4145–4152, https://doi.org/10.1007/s00146-024-02171-z.

Kabir, A., Shah, S., Haddad, A. and Raper, D.M.S., 'Introducing Our Custom GPT: An Example of the Potential Impact of Personalized GPT Builders on Scientific Writing', *World Neurosurgery*, No. 193, 2025, pp. 461–468, https://doi.org/10.1016/j.wneu.2024.10.041.

Kadhim, I. and Abed, F.M., 'The Potential of LiDAR and UAV-Photogrammetric Data Analysis to Interpret Archaeological Sites: A Case Study of Chun Castle in South-West England', *ISPRS International Journal of Geo-Information*, No. 10, Multidisciplinary Digital Publishing Institute, 2021, p. 41, https://doi.org/10.3390/ijgi10010041.

Kajiwara, Y. and Kawabata, K., 'AI Literacy for Ethical Use of Chatbot: Will Students Accept AI Ethics?', *Computers and Education: Artificial Intelligence*, No. 6, 2024, https://doi.org/10.1016/j.caeai.2024.100251.

Kaminski, K., Ludwiczak, J., Pawlicki, K., Alva, V. and Dunin-Horkawicz, S., 'pLM-BLAST: Distant Homology Detection Based on Direct Comparison of Sequence Representations from Protein Language Models', *Bioinformatics*, No. 39, 2023, https://doi.org/10.1093/bioinformatics/btad579.

Kao, Y.-S. and Yang, J., 'Deep Learning-Based Auto-Segmentation of Lung Tumor PET/CT Scans: A Systematic Review', *Clinical and Translational Imaging*, No. 10, 2022, pp. 217–223, https://doi.org/10.1007/s40336-022-00482-z.

Karagiorgi, G., Kasieczka, G., Kravitz, S., Nachman, B. and Shih, D., 'Machine Learning in the Search for New Fundamental Physics', *Nature Reviews Physics*, No. 4, Nature Publishing Group, 2022, pp. 399–412, https://doi.org/10.1038/s42254-022-00455-1.

Kim, J., Chen, M.L., Rezaei, S.J., Ramirez-Posada, M., Caswell-Jin, J.L. et al., 'Can Artificial Intelligence Generate Quality Research Topics Reflecting Patient Concerns?', 2024.

Kim, K., Lee, J., Yuen, R.S.H., Hannuksela, O.A. and Li, T.G.F., 'Identification of Lensed Gravitational Waves with Deep Learning', *The Astrophysical Journal*, No. 915, The American Astronomical Society, 2021, p. 119, https://doi.org/10.3847/1538-4357/ac0143.

Kim, S., Noh, J., Gu, G.H., Aspuru-Guzik, A. and Jung, Y., 'Generative Adversarial Networks for Crystal Structure Prediction', *ACS Central Science*, No. 6, American Chemical Society, 2020, pp. 1412–1420, https://doi.org/10.1021/acscentsci.0c00426.

Kirfel, L., J. MacCoun, R., Icard, T. and Gerstenberg, T., 'When AI Meets Counterfactuals: The Ethical Implications of Counterfactual World Simulation Models', *AI and Ethics*, 2025, https://doi.org/10.1007/s43681-025-00718-4.

Kirklin, S., Saal, J.E., Meredig, B., Thompson, A., Doak, J.W. et al., 'The Open Quantum Materials Database (OQMD): Assessing the Accuracy of DFT Formation Energies', *npj Computational Materials*, No. 1, Nature Publishing Group, 2015, https://doi.org/10.1038/npjcompumats.2015.10.

Knorr-Cetina, K., 'Epistemic cultures: How the sciences make knowledge', Harvard University Press, 1999.

Kobak, D., González-Márquez, R., Horvát, E.-Á. and Lause, J., 'Delving into LLM-Assisted Writing in Biomedical Publications through Excess Vocabulary', *Science Advances*, No. 11, American Association for the Advancement of Science, 2025, https://doi.org/10.1126/sciadv.adt3813.

Kodi Ramanah, D., Arendse, N. and Wojtak, R., 'AI-Driven Spatio-Temporal Engine for Finding Gravitationally Lensed Type Ia Supernovae', *Monthly Notices of the Royal Astronomical Society*, No. 512, 2022, pp. 5404–5417, https://doi.org/10.1093/mnras/stac838.

Kohnke, L., Zou, D., Ou, A.W. and Gu, M.M., 'Preparing Future Educators for AI-Enhanced Classrooms: Insights into AI Literacy and Integration', *Computers and Education: Artificial Intelligence*, No. 8, 2025, https://doi.org/10.1016/j.caeai.2025.100398.

Kovalevskiy, O., Mateos-Garcia, J. and Tunyasuvunakool, K., 'AlphaFold Two Years on: Validation and Impact', *Proceedings of the National Academy of Sciences*, No. 121, 2024, https://doi.org/10.1073/pnas.2315002121.

Krenn, M., Buffoni, L., Coutinho, B., Eppel, S., Foster, J.G. et al., 'Forecasting the Future of Artificial

Intelligence with Machine Learning-Based Link Prediction in an Exponentially Growing Knowledge Network', *Nature Machine Intelligence*, No. 5, Nature Publishing Group, 2023, pp. 1326–1335, https://doi.org/10.1038/s42256-023-00735-0.

Krenn, M., Malik, M., Fickler, R., Lapkiewicz, R. and Zeilinger, A., 'Automated Search for New Quantum Experiments', *Physical Review Letters*, No. 116, American Physical Society, 2016, https://doi.org/10.1103/PhysRevLett.116.090405.

Krenn, M., Pollice, R., Guo, S.Y., Aldeghi, M., Cervera-Lierta, A. et al., 'On Scientific Understanding with Artificial Intelligence', *Nature Reviews Physics*, No. 4, Nature Publishing Group, 2022, pp. 761–769, https://doi.org/10.1038/s42254-022-00518-3.

Krishnan, S., Crosby, C., Nandigam, V., Phan, M., Cowart, C. et al., 'OpenTopography: A Services Oriented Architecture for Community Access to LIDAR Topography', in: *Proceedings of the 2nd International Conference on Computing for Geospatial Research & Applications*, COM.Geo '11, Association for Computing Machinery, New York, NY, USA, 2011, pp. 1–8, https://doi.org/10.1145/1999320.1999327.

Kryshtafovych, A., Schwede, T., Topf, M., Fidelis, K. and Moult, J., 'Critical Assessment of Methods of Protein Structure Prediction (CASP)—Round XV', *Proteins: Structure, Function, and Bioinformatics*, No. 91, 2023, pp. 1539–1549, https://doi.org/10.1002/prot.26617.

Kryshtafovych, A., Schwede, T., Topf, M., Fidelis, K. and Moult, J., 'Critical Assessment of Methods of Protein Structure Prediction (CASP)—Round XIV', *Proteins: Structure, Function, and Bioinformatics*, No. 89, 2021, pp. 1607–1617, https://doi.org/10.1002/prot.26237.

Kücking, F., Hübner, U., Przysucha, M., Hannemann, N., Kutza, J.-O. et al., 'Automation Bias in AI-Decision Support: Results from an Empirical Study', *Studies in Health Technology and Informatics*, No. 317, 2024, pp. 298–304, https://doi.org/10.3233/SHTI240871.

Kuhn, T.S., 'The structure of scientific revolutions', University of Chicago Press, 1962.

Kumar, J.R.R., Chaturvedi, M., ANandhasilambarasan, D., Ashok, V., Nawadkar, A.R. et al., 'Assessing the Efficiency of AI for Multivariate Data Analysis', in: *2023 3rd International Conference on Smart Generation Computing, Communication and Networking (SMART GENCON)*, 2023, pp. 1–6, https://doi.org/10.1109/SMARTGENCON60755.2023.10442527.

Kumar, K., 'AI-Powered Citizen Science: Crowd-Learning, Data Verification, and Public Engagement', Social Science Research Network, Rochester, NY, 2025, https://doi.org/10.2139/ssrn.5251192.

Kumar, S., Datta, S., Singh, V., Singh, S.K. and Sharma, R., 'Opportunities and Challenges in Data-Centric AI', *IEEE Access*, No. 12, 2024, pp. 33173–33189, https://doi.org/10.1109/ACCESS.2024.3369417.

Lahav, D., Falcon, J.S., Kuehl, B., Johnson, S., Parasa, S. et al., 'A Search Engine for Discovery of Scientific Challenges and Directions', *Proceedings of the AAAI Conference on Artificial Intelligence*, No. 36, 2022, pp. 11982–11990, https://doi.org/10.1609/aaai.v36i11.21456.

Lai, Y.S., Neill, D., Płoskoń, M. and Ringer, F., 'Explainable Machine Learning of the Underlying Physics of High-Energy Particle Collisions', *Physics Letters B*, No. 829, 2022, https://doi.org/10.1016/j.physletb.2022.137055.

Lample, G. and Charton, F., 'Deep Learning for Symbolic Mathematics', arXiv, 2019, https://doi.org/10.48550/arXiv.1912.01412.

Lapp, E.C. and Lapp, L.W.P., 'Evaluating ChatGPT as a Viable Research Tool for Typological Investigations of Cultural Heritage Artefacts—Roman Clay Oil Lamps', *Archaeometry*, No. 66, 2024, pp. 696–717, https://doi.org/10.1111/arcm.12937.

Lathabai, H.H., Nandy, A. and Singh, V.K., ‘Institutional Collaboration Recommendation: An Expertise-Based Framework Using NLP and Network Analysis’, *Expert Systems with Applications*, No. 209, 2022, https://doi.org/10.1016/j.eswa.2022.118317.

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A.-L. et al., ‘Computational Social Science’, *Science*, No. 323, American Association for the Advancement of Science, 2009, pp. 721–723, https://doi.org/10.1126/science.1167742.

Ledford, H., ‘Consortium Solves Its 1,000th Protein Structure’, *Nature*, Nature Publishing Group, 2010, https://doi.org/10.1038/news.2010.500.

Lehr, S.A., Caliskan, A., Liyanage, S. and Banaji, M.R., ‘ChatGPT as Research Scientist: Probing GPT’s Capabilities as a Research Librarian, Research Ethicist, Data Generator, and Data Predictor’, *Proceedings of the National Academy of Sciences*, No. 121, 2024, https://doi.org/10.1073/pnas.2404328121.

Li, J., Lim, K., Yang, H., Ren, Z., Raghavan, S. et al., ‘AI Applications through the Whole Life Cycle of Material Discovery’, *Matter*, No. 3, 2020, pp. 393–432, https://doi.org/10.1016/j.matt.2020.06.011.

Li, L., Wang, G. and Freeman, G., ‘“It’s about Research. It’s Not about Language”: Understanding and Designing for Mitigating Non-Native English-Speaking Presenters’ Challenges in Live Q&A Sessions at Academic Conferences’, in: *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI ’25, Association for Computing Machinery, New York, NY, USA, 2025, pp. 1–15, https://doi.org/10.1145/3706598.3713124.

Li, L., Xu, W., Guo, J., Zhao, R., Li, X. et al., ‘Chain of Ideas: Revolutionizing Research Via Novel Idea Development with LLM Agents’, arXiv, 2024.

Li, R., Patel, T., Wang, Q. and Du, X., ‘MLR-Copilot: Autonomous Machine Learning Research Based on Large Language Models Agents’, arXiv, 2024, https://doi.org/10.48550/arXiv.2408.14033.

Li, X., Zhang, L., Yang, J. and Teng, F., ‘Role of Artificial Intelligence in Medical Image Analysis: A Review of Current Trends and Future Directions’, *Journal of Medical and Biological Engineering*, No. 44, 2024, pp. 231–243, https://doi.org/10.1007/s40846-024-00863-x.

Liang, T., Duarte, N. and Yue, G.X.-G., ‘An Evolutionary Study of the Impact of Artificial Intelligence Technology on Foreign Language Education’, *International Journal of Emerging Technologies in Learning (iJET)*, No. 18, 2023, pp. 190–204, https://doi.org/10.3991/ijet.v18i19.43821.

Liang, W., Zhang, Y., Wu, Z., Lepp, H., Ji, W. et al., ‘Mapping the Increasing Use of LLMs in Scientific Papers’, in: *First Conference on Language Modeling*, 2024.

Lim, L.X.J. and Qiu, Z., ‘Mu-Net: ConvNext-Based U-Nets for Cosmic Muon Tomography’, https://doi.org/10.48550/arXiv.2312.17265.

Lin, H. and Zhang, Y., ‘The Risks of Using Large Language Models for Text Annotation in Social Science Research’, arXiv, 2025, https://doi.org/10.48550/arXiv.2503.22040.

Lin, P.-Y., Huang, S.-C., Chen, K.-L., Huang, Y.-C., Liao, C.-Y. et al., ‘Analysing Protein Complexes in Plant Science: Insights and Limitation with AlphaFold 3’, *Botanical Studies*, No. 66, 2025, p. 14, https://doi.org/10.1186/s40529-025-00462-2.

Lin, Z., ‘Techniques for Supercharging Academic Writing with Generative AI’, *Nature Biomedical Engineering*, No. 9, Nature Publishing Group, 2025, pp. 426–431, https://doi.org/10.1038/s41551-024-01185-8.

Lin, Z., ‘Progress and Challenges in the Symbiosis of AI with Science and Medicine’, *European Journal of Clinical Investigation*, No. 54, 2024, https://doi.org/10.1111/eci.14222.

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z. et al., ‘Evolutionary-Scale Prediction of Atomic-Level Protein Structure with a Language Model’, *Science*, No. 379, American Association for the Advancement of Science, 2023, pp. 1123–1130, https://doi.org/10.1126/science.ade2574.

Lipkova, J., Chen, R.J., Chen, B., Lu, M.Y., Barbieri, M. et al., 'Artificial Intelligence for Multimodal Data Integration in Oncology', *Cancer Cell*, No. 40, Elsevier, 2022, pp. 1095–1110, https://doi.org/10.1016/j.ccell.2022.09.012.

Liu, C., Han, Y., Xu, H., Yang, S., Wang, K. et al., 'A Community Detection and Graph-Neural-Network-Based Link Prediction Approach for Scientific Literature', *Mathematics*, No. 12, Multidisciplinary Digital Publishing Institute, 2024, p. 369, https://doi.org/10.3390/math12030369.

Liu, H., Zhou, Y., Li, M., Yuan, C. and Tan, C., 'Literature Meets Data: A Synergistic Approach to Hypothesis Generation', arXiv, 2025, https://doi.org/10.48550/arXiv.2410.17309.

Liu, J.Q.J., Hui, K.T.K., Al Zoubi, F., Zhou, Z.Z.X., Samartzis, D. et al., 'The Great Detectives: Humans versus AI Detectors in Catching Large Language Model-Generated Medical Writing', *International Journal for Educational Integrity*, No. 20, BioMed Central, 2024, pp. 1–14, https://doi.org/10.1007/s40979-024-00155-6.

Liu, S., Lu, Yizhou, Chen, S., Hu, X., Zhao, J. et al., 'DrugAgent: Automating AI-Aided Drug Discovery Programming through LLM Multi-Agent Collaboration', arXiv, 2025, https://doi.org/10.48550/arXiv.2411.15692.

Liu, S.-C., Wang, S., Chang, T., Lin, W., Hsiung, C.-W. et al., 'JarviX: A LLM No Code Platform for Tabular Data Analysis and Optimization', in: Wang, M., and Zitouni, I. (Eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, Association for Computational Linguistics, Singapore, 2023, pp. 622–630, https://doi.org/10.18653/v1/2023.emnlp-industry.59.

Liu, X., Wu, Z., Wu, X., Lu, P., Chang, K.-W. et al., 'Are LLMs Capable of Data-Based Statistical and Causal Reasoning? Benchmarking Advanced Quantitative Reasoning with Data', in: *Findings of the Association for Computational Linguistics: ACL 2024*, Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 9215–9235, https://doi.org/10.18653/v1/2024.findings-acl.548.

Liu, Yuhan, Shah, A., Ackerman, J. and Saha, M., 'Exploring the Design Space of Real-Time LLM Knowledge Support Systems: A Case Study of Jargon Explanations', in: *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, Association for Computing Machinery, New York, NY, USA, 2025, pp. 1–20, https://doi.org/10.1145/3706598.3714262.

Liu, Yiren, Sharma, P., Oswal, M., Xia, H. and Huang, Y., 'PersonaFlow: Designing LLM-Simulated Expert Perspectives for Enhanced Research Ideation', in: *Proceedings of the 2025 ACM Designing Interactive Systems Conference*, DIS '25, Association for Computing Machinery, New York, NY, USA, 2025, pp. 506–534, https://doi.org/10.1145/3715336.3735789.

Liu, Z., Liu, K., Zhu, Y., Lei, X., Yang, Z. et al., 'AIGS: Generating Science from AI-Powered Automated Falsification', arXiv, 2024, https://doi.org/10.48550/arXiv.2411.11910.

Lo, K., Chang, J.C., Head, A., Bragg, J., Zhang, A.X. et al., 'The Semantic Reader Project: Augmenting Scholarly Documents through AI-Powered Interactive Reading Interfaces', arXiv, 2023, https://doi.org/10.48550/arXiv.2303.14334.

Long, C.X. and Wang, J., 'China's Patent Promotion Policies and Its Quality Implications', *Science and Public Policy*, No. 46, 2019, pp. 91–104, https://doi.org/10.1093/scipol/scy040.

Long, T., Fortunato, N.M., Opahle, I., Zhang, Y., Samathrakis, I. et al., 'Constrained Crystals Deep Convolutional Generative Adversarial Network for the Inverse Design of Crystal Structures', *npj Computational Materials*, No. 7, Nature Publishing Group, 2021, p. 66, https://doi.org/10.1038/s41524-021-00526-4.

Longino, H.E., 'The fate of knowledge', Princeton University Press, 2002.

Lu, C., 'Rethinking Artificial Intelligence from the Perspective of Interdisciplinary Knowledge Production', *AI & SOCIETY*, No. 39, 2024, pp. 3059–3060, https://doi.org/10.1007/s00146-023-01839-2.

Lund, B. and Shamsi, A., 'Examining the Use of Supportive and Contrasting Citations in Different Disciplines: A Brief Study Using Scite (Scite.Ai) Data', *Scientometrics*, No. 128, 2023, pp. 4895–4900, https://doi.org/10.1007/s11192-023-04781-8.

Luo, Z., Yang, Z., Xu, Z., Yang, W. and Du, X., 'LLM4SR: A Survey on Large Language Models for Scientific Research', arXiv, 2025, https://doi.org/10.48550/arXiv.2501.04306.

Maddigan, P. and Susnjak, T., 'Chat2VIS: Generating Data Visualizations via Natural Language Using ChatGPT, Codex and GPT-3 Large Language Models', *IEEE Access*, No. 11, 2023, pp. 45181–45193, https://doi.org/10.1109/ACCESS.2023.3274199.

Majumder, B.P., Surana, H., Agarwal, D., Mishra, B.D., Meena, A. et al., 'DiscoveryBench: Towards Data-Driven Discovery with Large Language Models', arXiv, 2024, https://doi.org/10.48550/arXiv.2407.01725.

Makke, N. and Chawla, S., 'Interpretable Scientific Discovery with Symbolic Regression: A Review', *Artificial Intelligence Review*, No. 57, 2024, p. 2, https://doi.org/10.1007/s10462-023-10622-0.

Manikandan, S., Kaviya, R.S., Shreeharan, D.H., Subbaiya, R., Vickram, S. et al., 'Artificial Intelligence-Driven Sustainability: Enhancing Carbon Capture for Sustainable Development Goals– A Review', *Sustainable Development*, No. 33, 2025, pp. 2004–2029, https://doi.org/10.1002/sd.3222.

Markowitz, D.M., 'From Complexity to Clarity: How AI Enhances Perceptions of Scientists and the Public's Understanding of Science', *PNAS Nexus*, No. 3, 2024, https://doi.org/10.1093/pnasnexus/pgae387.

Martin, A.C., Orengo, C.A., Hutchinson, E.G., Jones, S., Karmirantzou, M. et al., 'Protein Folds and Functions', *Structure*, No. 6, Elsevier, 1998, pp. 875–884, https://doi.org/10.1016/S0969-2126(98)00089-6.

Matsui, Y., Gidding, A., Levy, T.E., Kuester, F. and DeFanti, T.A., 'ArchaeoSTOR Map: Publishing Archaeological Geodata on the Web', in: *Proceedings of the 3rd International Conference on Computing for Geospatial Research and Applications*, COM.Geo '12, Association for Computing Machinery, New York, NY, USA, 2012, pp. 1–2, https://doi.org/10.1145/2345316.2345355.

McCoy, M.D., 'Geospatial Big Data and Archaeology: Prospects and Problems Too Great to Ignore', *Journal of Archaeological Science*Archaeological GIS Today: Persistent Challenges, Pushing Old Boundaries, and Exploring New Horizons, No. 84, 2017, pp. 74–94, https://doi.org/10.1016/j.jas.2017.06.003.

McDonald, N., Johri, A., Ali, A. and Collier, A.H., 'Generative Artificial Intelligence in Higher Education: Evidence from an Analysis of Institutional Policies and Guidelines', *Computers in Human Behavior: Artificial Humans*, No. 3, 2025, https://doi.org/10.1016/j.chbah.2025.100121.

McGinness, L., Baumgartner, P., Onyango, E. and Lema, Z., 'Highlighting Case Studies in LLM Literature Review of Interdisciplinary System Science', in: Gong, M., Song, Y., Koh, Y.S., Xiang, W., and Wang, D. (Eds.), *AI 2024: Advances in Artificial Intelligence*, Lecture Notes in Computer Science, Springer Nature Singapore, Singapore, 2024, pp. 29–43, https://doi.org/10.1007/978-981-96-0348-0_3.

Mece, E.K., Paci, H. and Binjaku, K., 'The Application Of Machine Learning In Test Case Prioritization - A Review', *European Journal of Electrical Engineering and Computer Science*, No. 4, 2020, https://doi.org/10.24018/ejece.2020.4.1.128.

Medina-Ortiz, D., Khalifeh, A., Anvari-Kazemabad, H. and Davari, M.D., 'Interpretable and Explainable

Predictive Machine Learning Models for Data-Driven Protein Engineering', *Biotechnology Advances*, No. 79, 2025, https://doi.org/10.1016/j.biotechadv.2024.108495.

Melliti, M., 'AI in MA Thesis Writing: The Use of Lexical Patterns to Study the ChatGPT Influence', *International Journal of TESOL Studies*, 2024, https://doi.org/10.58304/ijts.20240305.

Mengaldo, G., 'Explain the Black Box for the Sake of Science: The Scientific Method in the Era of Generative Artificial Intelligence', arXiv, 2025, https://doi.org/10.48550/arXiv.2406.10557.

Merchant, A., Batzner, S., Schoenholz, S.S., Aykol, M., Cheon, G. et al., 'Scaling Deep Learning for Materials Discovery', *Nature*, No. 624, Nature Publishing Group UK London, 2023, pp. 80–85.

Messeri, L. and Crockett, M.J., 'Artificial Intelligence and Illusions of Understanding in Scientific Research', *Nature*, No. 627, Nature Publishing Group, 2024, pp. 49–58, https://doi.org/10.1038/s41586-024-07146-0.

Miah, A.S.M., Tusher, M.M.R., Hossain, Md Moazzem, Hossain, Md Mamun, Rahim, M.A. et al., 'ChatGPT in Research and Education: Exploring Benefits and Threats', arXiv, 2024.

Millar, F., 'Epigraphy', in: Crawford, M. (Ed.), *Sources for Ancient History*, Sources of History, Cambridge University Press, Cambridge, 1983, pp. 80–136, https://doi.org/10.1017/CBO9780511622229.003.

Mirdita, M., von den Driesch, L., Galiez, C., Martin, M.J., Söding, J. et al., 'Uniclust Databases of Clustered and Deeply Annotated Protein Sequences and Alignments', *Nucleic Acids Research*, No. 45, 2017, pp. 170–176, https://doi.org/10.1093/nar/gkw1081.

Misra, K. and Kim, N., 'Generating Novel Experimental Hypotheses from Language Models: A Case Study on Cross-Dative Generalization', arXiv, 2024, https://doi.org/10.48550/arXiv.2408.05086.

MIT Economics, 'Assuring an Accurate Research Record', 2025.

Mohlehli, M.G., Adam, E. and Schoeman, M.H., 'The Potential for Lidar Using Support Vector Machine (Svm) to Detect Archaeological Stone-Walled Structures in Khutwaneng, Bokoni', *The South African Archaeological Bulletin*, No. 78, South African Archaeological Society, 2023, pp. 33–42.

Mollaki, V., 'Death of a Reviewer or Death of Peer Review Integrity? The Challenges of Using AI Tools in Peer Reviewing and the Need to Go beyond Publishing Policies', *Research Ethics*, No. 20, SAGE Publications Ltd, 2024, pp. 239–250, https://doi.org/10.1177/17470161231224552.

Monge Roffarello, A., Calò, T., Scibetta, L. and De Russis, L., 'Investigating How Computer Science Researchers Design Their Co-Writing Experiences With AI', in: *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, ACM, Yokohama Japan, 2025, pp. 1–17, https://doi.org/10.1145/3706598.3713205.

Montgomery, L.M. and Fryer, T.C., 'The Future of Archaeology Is (Still) Community Collaboration', *Antiquity*, No. 97, 2023, pp. 795–809, https://doi.org/10.15184/aqy.2023.98.

Mostafapour, M., Fortier, J.H., Pacheco, K., Murray, H. and Garber, G., 'Evaluating Literature Reviews Conducted by Humans Versus ChatGPT: Comparative Study', *JMIR AI*, No. 3, 2024, https://doi.org/10.2196/56537.

Moult, J., Pedersen, J.T., Judson, R. and Fidelis, K., 'A Large-Scale Experiment to Assess Protein Structure Prediction Methods', *Proteins: Structure, Function, and Bioinformatics*, No. 23, 1995, pp. 2-4, https://doi.org/10.1002/prot.340230303.

Mousavi, S.M. and Beroza, G.C., 'Machine Learning in Earthquake Seismology', *Annual Review of Earth and Planetary Sciences*, No. 51, Annual Reviews, 2023, pp. 105–129, https://doi.org/10.1146/annurev-earth-071822-100323.

Muñoz-Ortiz, A., Gómez-Rodríguez, C. and Vilares, D., 'Contrasting Linguistic Patterns in Human and LLM-Generated News Text', *Artificial Intelligence Review*, No. 57, 2024, p. 265, https://doi.org/10.1007/s10462-024-10903-2.

Musslick, S., Bartlett, L.K., Chandramouli, S.H., Dubova, M., Gobet, F. et al., 'Automating the Practice of Science: Opportunities, Challenges, and Implications', *Proceedings of the National Academy of Sciences*, No. 122, 2025, https://doi.org/10.1073/pnas.2401238121.

Narayanan, A. and Kapoor, S., 'Why an Overreliance on AI-Driven Modelling Is Bad for Science', *Nature*, 2025.

National Research Council, 'A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas', National Academies Press, Washington, D.C., 2012, https://doi.org/10.17226/13165.

Nature, 'Metascience Can Improve Science — but It Must Be Useful to Society, Too', *Nature Editorials*, URL https://www.nature.com/articles/d41586-025-02065-0.

Nicholas, D., Swigon, M., Clark, D., Abrizah, A., Revez, J. et al., 'The Impact of Generative AI on the Scholarly Communications of Early Career Researchers: An International, Multi-Disciplinary Study', *Learned Publishing*, No. 37, 2024, https://doi.org/10.1002/leap.1628.

Nigam, H., Patwardhan, M., Vig, L. and Shroff, G., 'An Interactive Co-Pilot for Accelerated Research Ideation', in: Blodgett, S.L., Cercas Curry, A., Dev, S., Madaio, M., Nenkova, A. et al. (Eds.), *Proceedings of the Third Workshop on Bridging Human–Computer Interaction and Natural Language Processing*, Association for Computational Linguistics, Mexico City, Mexico, 2024, pp. 60–73, https://doi.org/10.18653/v1/2024.hcinlp-1.6.

NIST, 'REMI: REsource for Materials Informatics - A Knowledge Sharing Platform'.

Noh, J., Kim, J., Stein, H.S., Sanchez-Lengeling, B., Gregoire, J.M. et al., 'Inverse Design of Solid-State Materials via a Continuous Representation', *Matter*, No. 1, 2019, pp. 1370–1384, https://doi.org/10.1016/j.matt.2019.08.017.

Nouira, A., Sokolovska, N. and Crivello, J.-C., 'CrystalGAN: Learning to Discover Crystallographic Structures with Generative Adversarial Networks', in: *Proceedings of the AAAI 2019 Spring Symposium on Combining Machine Learning with Knowledge Engineering*, CEUR Workshop Proceedings, Palo Alto, California, USA, 2019.

OECD, 'Explanatory Memorandum on the Updated OECD Definition of an AI System' (OECD Artificial Intelligence Papers No. 8),OECD Artificial Intelligence Papers,2024, https://doi.org/10.1787/623da898-en.

OECD, 'Artificial Intelligence in Science. Challenges, Opportunities and the Future of Research',2023.

OECD, 'OECD Patent Statistics Manual', Paris, 2009.

Oelen, A., Jaradeh, M.Y. and Auer, S., 'ORKG ASK: A Neuro-Symbolic Scholarly Search and Exploration System', arXiv, 2024, https://doi.org/10.48550/arXiv.2412.04977.

Oksanen, J., 'Bridging the Integrity Gap: Towards AI-Assisted Design Research', in: *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, CHI EA '24, Association for Computing Machinery, New York, NY, USA, 2024, pp. 1–5, https://doi.org/10.1145/3613905.3647962.

Onishi, Y., 'Deep Learning-Aided Research and the Aim-of-Science Controversy', *Journal for General Philosophy of Science*, No. 56, 2025, pp. 77–95, https://doi.org/10.1007/s10838-023-09667-0.

Orengo, H.A., Conesa, F.C., Garcia-Molsosa, A., Lobo, A., Green, A.S. et al., 'Automated Detection of Archaeological Mounds Using Machine-Learning Classification of Multisensor and Multitemporal Satellite Data', *Proceedings of the National*

Academy of Sciences, No. 117, 2020, pp. 18240–18250, https://doi.org/10.1073/pnas.2005583117.

Ormö, M., Cubitt, A.B., Kallio, K., Gross, L.A., Tsien, R.Y. et al., 'Crystal Structure of the Aequorea Victoria Green Fluorescent Protein', *Science*, No. 273, American Association for the Advancement of Science, 1996, pp. 1392–1395, https://doi.org/10.1126/science.273.5280.1392.

Oturak, S., 'The Collection of Database and Dataset Resources in Materials Science'.

Pagliaro, A. and Sangiorgi, P., 'AI in Experiments: Present Status and Future Prospects', *Applied Sciences*, No. 13, Multidisciplinary Digital Publishing Institute, 2023, https://doi.org/10.3390/app131810415.

Pal, S., 'A Paradigm Shift in Research: Exploring the Intersection of Artificial Intelligence and Research Methodology', *International Journal of Innovative Research in Engineering & Multidisciplinary Physical Sciences*, No. 11, 2023, https://doi.org/10.37082/IJIRMPS.v11.i3.230125.

Pang, S., Nol, E. and Heng, K., 'Generative AI as a Personal Tutor for English Language Learning: A Review of Benefits and Concerns', *International Journal of Changes in Education*, 2025, https://doi.org/10.47852/bonviewIJCE52023724.

Park, H., Li, Z. and Walsh, A., 'Has Generative Artificial Intelligence Solved Inverse Materials Design?', *Matter*, No. 7, Elsevier, 2024, pp. 2355–2367.

Park, K.-S. and Choi, H., 'How to Harness the Power of GPT for Scientific Research: A Comprehensive Review of Methodologies, Applications, and Ethical Considerations', *Nuclear Medicine and Molecular Imaging*, No. 58, 2024, pp. 323–331, https://doi.org/10.1007/s13139-024-00876-z.

Parsan, N., Yang, D.J. and Yang, J.J., 'Towards Interpretable Protein Structure Prediction with Sparse Autoencoders', arXiv, 2025, https://doi.org/10.48550/arXiv.2503.08764.

Passaro, S., Corso, G., Wohlwend, J., Reveiz, M., Thaler, S. et al., 'Boltz-2: Towards Accurate and Efficient Binding Affinity Prediction', bioRxiv, 2025, https://doi.org/10.1101/2025.06.14.659707.

Pataranutaporn, P., Powdthavee, N. and Maes, P., 'Can Ai Solve the Peer Review Crisis? A Large-Scale Experiment on Llm's Performance and Biases in Evaluating Economics Papers', Social Science Research Network, Rochester, NY, 2025.

Patel, K., 'Ethical Reflections On Data-Centric AI: Balancing Benefits And Risks', Social Science Research Network, Rochester, NY, 2024, https://doi.org/10.2139/ssrn.4993089.

Paysan-Lafosse, T., Andreeva, A., Blum, M., Chuguransky, S.R., Grego, T. et al., 'The Pfam Protein Families Database: Embracing AI/ML', *Nucleic Acids Research*, No. 53, 2025, pp. 523–534, https://doi.org/10.1093/nar/gkae997.

Peppin, A., Reuel, A., Casper, S., Jones, E., Strait, A. et al., 'The Reality of AI and Biorisk', arXiv, 2024, https://doi.org/10.48550/arXiv.2412.01946.

Persaud, D., Ward, L. and Hattrick-Simpers, J., 'Reproducibility in Materials Informatics: Lessons from "A General-Purpose Machine Learning Framework for Predicting Properties of Inorganic Materials"', *Digital Discovery*, No. 3, 2024, pp. 281–286, https://doi.org/10.1039/D3DD00199G.

Peters, U. and Chin-Yee, B., 'Generalization Bias in Large Language Model Summarization of Scientific Research', *Royal Society Open Science*, No. 12, 2025, https://doi.org/10.1098/rsos.241776.

Polakova, P. and Klimova, B., 'Using DeepL Translator in Learning English as an Applied Foreign Language – An Empirical Pilot Study', *Heliyon*, No. 9, Elsevier, 2023, https://doi.org/10.1016/j.heliyon.2023.e18595.

Popper, K., 'The logic of scientific discovery', Hutchinson, 1959.

Porta-Pardo, E., Ruiz-Serra, V., Valentini, S. and Valencia, A., 'The Structural Coverage of the

Human Proteome before and after AlphaFold', *PLOS Computational Biology*, No. 18, Public Library of Science, 2022, https://doi.org/10.1371/journal.pcbi.1009818.

Pretolesi, D., Stanzani, I., Ravera, S., Vian, A. and Barla, A., 'Artificial Intelligence and Network Science as Tools to Illustrate Academic Research Evolution in Interdisciplinary Fields: The Case of Italian Design', *PLOS ONE*, No. 20, Public Library of Science, 2025, https://doi.org/10.1371/journal.pone.0315216.

Pu, K., Feng, K.J.K., Grossman, T., Hope, T., Mishra, B.D. et al., 'IdeaSynth: Iterative Research Idea Development Through Evolving and Composing Idea Facets with Literature-Grounded Feedback', arXiv, 2024, https://doi.org/10.48550/arXiv.2410.04025.

Pudasaini, S., Miralles-Pechuán, L., Lillis, D. and Llorens Salvador, M., 'Survey on AI-Generated Plagiarism Detection: The Impact of Large Language Models on Academic Integrity', *Journal of Academic Ethics*, 2024, https://doi.org/10.1007/s10805-024-09576-x.

Pugliese, R., Badini, S., Frontoni, E. and Regondi, S., 'Generative Artificial Intelligence for Advancing Discovery and Design in Biomateriomics', *Intelligent Computing*, No. 4, American Association for the Advancement of Science, 2025, https://doi.org/10.34133/icomputing.0117.

Purificato, E., Boratto, L. and De Luca, E.W., 'Toward a Responsible Fairness Analysis: From Binary to Multiclass and Multigroup Assessment in Graph Neural Network-Based User Modeling Tasks', *Minds and Machines*, No. 34, 2024, p. 33, https://doi.org/10.1007/s11023-024-09685-x.

Pyzer-Knapp, E.O., Pitera, J.W., Staar, P.W.J., Takeda, S., Laino, T. et al., 'Accelerating Materials Discovery Using Artificial Intelligence, High Performance Computing and Robotics', *npj Computational Materials*, No. 8, 2022, p. 84, https://doi.org/10.1038/s41524-022-00765-z.

Raayoni, G., Gottlieb, S., Manor, Y., Pisha, G., Harris, Y. et al., 'Generating Conjectures on Fundamental Constants with the Ramanujan Machine', *Nature*, No. 590, Nature Publishing Group, 2021, pp. 67–73, https://doi.org/10.1038/s41586-021-03229-4.

Raghuwanshi, P., 'Revolutionizing Semicondutor Design and Manufacturing with AI', *Journal of Knowledge Learning and Science Technology ISSN: 2959-6386 (online)*, No. 3, 2024, pp. 272–277, https://doi.org/10.60087/jklst.vol3.n3.p.272-277.

Rajwal, S., Zhang, Z., Chen, Y., Rogers, H., Sarker, A. et al., 'Applications of Natural Language Processing and Large Language Models for Social Determinants of Health: Protocol for a Systematic Review', *JMIR Research Protocols*, No. 14, 2025, https://doi.org/10.2196/66094.

Rao, H., Leung, C. and Miao, C., 'Can ChatGPT Assess Human Personalities? A General Evaluation Framework', arXiv, 2023, https://doi.org/10.48550/arXiv.2303.01248.

Reddy, C.K. and Shojaee, P., 'Towards Scientific Discovery with Generative AI: Progress, Opportunities, and Challenges', *Proceedings of the AAAI Conference on Artificial Intelligence*, No. 39, 2025, pp. 28601–28609, https://doi.org/10.1609/aaai.v39i27.35084.

Reeve, S.T., Guzman, D.M., Alzate-Vargas, L., Haley, B., Liao, P. et al., 'Online Simulation Powered Learning Modules for Materials Science', *MRS Advances*, No. 4, 2019, pp. 2727–2742, https://doi.org/10.1557/adv.2019.287.

Ren, Z., Tian, S.I.P., Noh, J., Oviedo, F., Xing, G. et al., 'An Invertible Crystallographic Representation for General Inverse Design of Inorganic Crystals with Targeted Properties', *Matter*, No. 5, Elsevier, 2022, pp. 314–335, https://doi.org/10.1016/j.matt.2021.11.032.

Resce, G., Zinilli, A. and Cerulli, G., 'Machine Learning Prediction of Academic Collaboration Networks', *Scientific Reports*, No. 12, Nature Publishing Group, 2022, https://doi.org/10.1038/s41598-022-26531-1.

Resnik, D.B. and Hosseini, M., 'The Ethics of Using Artificial Intelligence in Scientific Research: New Guidance Needed for a New Tool', *AI and Ethics*, No. 5, 2025, pp. 1499–1521, https://doi.org/10.1007/s43681-024-00493-8.

Richardson, L., Allen, B., Baldi, G., Beracochea, M., Bileschi, M.L. et al., 'MGnify: The Microbiome Sequence Data Analysis Resource in 2023', *Nucleic Acids Research*, No. 51, 2023, pp. 753–759, https://doi.org/10.1093/nar/gkac1080.

Richards-Rissetto, H., Newton, D. and Al Zadjali, A., 'A 3D POINT CLOUD DEEP LEARNING APPROACH USING LIDAR TO IDENTIFY ANCIENT MAYA ARCHAEOLOGICAL SITES', *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, No. VIII-M-1–2021, 2021, pp. 133–139, https://doi.org/10.5194/isprs-annals-VIII-M-1-2021-133-2021.

Rietz, T. and Maedche, A., 'Cody: An AI-Based System to Semi-Automate Coding for Qualitative Research', in: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, Association for Computing Machinery, New York, NY, USA, 2021, pp. 1–14, https://doi.org/10.1145/3411764.3445591.

Righi, R., Lopez Cobo, M., Samoili, S., Cardona, M., Vazquez-Prada Baillet, M. et al., 'EU in the global Artificial Intelligence landscape', JRC125613, Publications Office of the European Union, Luxembourg, 2021.

Robert, L., 'Épigraphie', in: *L'Histoire et ses méthodes*, Gallimard, 1961, pp. 453–497, https://doi.org/10.3917/gallrel.sama.1961.01.0453.

Rolnik, Z., 'The Impact of Artificial Intelligence on Academic Research', *Universal Library of Innovative Research and Studies*, No. 01, 2024, pp. 9–11, https://doi.org/10.70315/uloap.ulirs.2024.0101002.

Roohani, Y.H., Hua, T.J., Tung, P.-Y., Bounds, L.R., Yu, F.B. et al., 'Virtual Cell Challenge: Toward a Turing Test for the Virtual Cell', *Cell*, No. 188, Elsevier, 2025, pp. 3370–3374, https://doi.org/10.1016/j.cell.2025.06.008.

Rosala, M. and Moran, K., 'Synthetic Users: If, When, and How to Use AI-Generated "Research"', *Nielsen Norman Group*, 2024, https://www.nngroup.com/articles/synthetic-users.

Rosignoli, S., Pacelli, M., Manganiello, F. and Paiardini, A., 'An Outlook on Structural Biology after AlphaFold: Tools, Limits and Perspectives', *FEBS Open Bio*, No. 15, 2025, pp. 202–222, https://doi.org/10.1002/2211-5463.13902.

Ruediger, D., McCracken, C. and Skinner, M., 'Adoption of Generative AI by Academic Biomedical Researchers', Ithaka S+R, 2024, https://doi.org/10.18665/sr.321415.

Ruffolo, J.A., Chu, L.-S., Mahajan, S.P. and Gray, J.J., 'Fast, Accurate Antibody Structure Prediction from Deep Learning on Massive Set of Natural Antibodies', *Nature Communications*, No. 14, 2023, p. 2389, https://doi.org/10.1038/s41467-023-38063-x.

Ryan, M., 'In AI We Trust: Ethics, Artificial Intelligence, and Reliability', *Science and Engineering Ethics*, No. 26, 2020, pp. 2749–2767, https://doi.org/10.1007/s11948-020-00228-y.

Sá, R., Gonçalves, L.J., Medina, J., Neves, A., Marsh, F. et al., 'Odyssey: A Spatial Data Infrastructure for Archaeology', *Journal of Computer Applications in Archaeology*, No. 7, 2024, https://doi.org/10.5334/jcaa.115.

Sadybekov, A.V. and Katritch, V., 'Computational Approaches Streamlining Drug Discovery', *Nature*, No. 616, Nature Publishing Group, 2023, pp. 673–685, https://doi.org/10.1038/s41586-023-05905-z.

Saeidnia, H.R., Hosseini, E., Abdoli, S. and Ausloos, M., 'Unleashing the Power of AI: A Systematic Review of Cutting-Edge Techniques in AI-Enhanced Scientometrics, Webometrics and Bibliometrics', *Library Hi Tech*, No. ahead-of-print, Emerald Publishing Limited, 2024, https://doi.org/10.1108/LHT-10-2023-0514.

Saeidnia, H.R., Hosseini, E., Lund, B., Tehrani, M.A., Zaker, S. et al., 'Artificial Intelligence in the Battle against Disinformation and Misinformation: A Systematic Review of Challenges and Approaches', *Knowledge and Information Systems*, No. 67, 2025, pp. 3139–3158, https://doi.org/10.1007/s10115-024-02337-7.

Sakai, M., Sakurai, A., Lu, S., Olano, J., Albrecht, C.M. et al., 'AI-Accelerated Nazca Survey Nearly Doubles the Number of Known Figurative Geoglyphs and Sheds Light on Their Purpose', *Proceedings of the National Academy of Sciences*, No. 121, 2024, https://doi.org/10.1073/pnas.2407652121.

Salcedo-Sanz, S., Pérez-Aracil, J., Ascenso, G., Del Ser, J., Casillas-Pérez, D. et al., 'Analysis, Characterization, Prediction, and Attribution of Extreme Atmospheric Events with Machine Learning and Deep Learning Techniques: A Review', *Theoretical and Applied Climatology*, No. 155, 2024, pp. 1–44, https://doi.org/10.1007/s00704-023-04571-5.

Salvagno, M., Taccone, F.S. and Gerli, A.G., 'Can Artificial Intelligence Help for Scientific Writing?', *Critical Care*, No. 27, 2023, p. 75, https://doi.org/10.1186/s13054-023-04380-2.

Sami, A.M., Rasheed, Z., Kemell, K.-K., Waseem, M., Kilamo, T. et al., 'System for Systematic Literature Review Using Multiple AI Agents: Concept and an Empirical Evaluation', arXiv, 2024, https://doi.org/10.48550/arXiv.2403.08399.

Sarker, I.H., 'AI-Based Modeling: Techniques, Applications and Research Issues Towards Automation, Intelligent and Smart Systems', *SN Computer Science*, No. 3, 2022, p. 158, https://doi.org/10.1007/s42979-022-01043-x.

Scheidgen, M., Himanen, L., Ladines, A.N., Sikter, D., Nakhaee, M. et al., 'NOMAD: A Distributed Web-Based Platform for Managing Materials Science Research Data', *Journal of Open Source Software*, No. 8, 2023, p. 5388, https://doi.org/10.21105/joss.05388.

Schemmer, M., Kuehl, N., Benz, C., Bartos, A. and Satzger, G., 'Appropriate Reliance on AI Advice: Conceptualization and the Effect of Explanations', in: *Proceedings of the 28th International Conference on Intelligent User Interfaces*, IUI '23, Association for Computing Machinery, New York, NY, USA, 2023, pp. 410–422, https://doi.org/10.1145/3581641.3584066.

Schetinger, V., Di Bartolomeo, S., El-Assady, M., McNutt, A., Miller, M. et al., 'Doom or Deliciousness: Challenges and Opportunities for Visualization in the Age of Generative Models', *Computer Graphics Forum*, No. 42, 2023, pp. 423–435, https://doi.org/10.1111/cgf.14841.

Schmidgall, S., Su, Y., Wang, Z., Sun, X., Wu, J. et al., 'Agent Laboratory: Using LLM Agents as Research Assistants', arXiv, 2025, https://doi.org/10.48550/ARXIV.2501.04227.

Schmidt, J., Hoffmann, N., Wang, H.-C., Borlido, P., Carriço, P.J.M.A. et al., 'Machine-Learning-Assisted Determination of the Global Zero-Temperature Phase Diagram of Materials', *Advanced Materials*, No. 35, 2023, https://doi.org/10.1002/adma.202210788.

Schölkopf, B., 'Causality for Machine Learning', in: *Probabilistic and Causal Inference: The Works of Judea Pearl*, Association for Computing Machinery, New York, NY, USA, 2022, pp. 765–804.

Schryen, G., Marrone, M. and Yang, J., 'Exploring the Scope of Generative AI in Literature Review Development', *Electronic Markets*, No. 35, 2025, p. 13, https://doi.org/10.1007/s12525-025-00754-2.

Seghier, M.L., 'AI-Powered Peer Review Needs Human Supervision', *Journal of Information, Communication and Ethics in Society*, No. 23, 2025, pp. 104–116, https://doi.org/10.1108/JICES-09-2024-0132.

Serey, J., Alfaro, M., Fuertes, G., Vargas, M., Durán, C. et al., 'Pattern Recognition and Deep Learning Technologies, Enablers of Industry 4.0, and Their Role in Engineering Research', *Symmetry*, No. 15, Multidisciplinary Digital Publishing Institute, 2023, p. 535, https://doi.org/10.3390/sym15020535.

Shawki, N., Nunez, R.R., Obeid, I. and Picone, J., 'On Automating Hyperparameter Optimization for Deep Learning Applications', in: *2021 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*, 2021, pp. 1–7, https://doi.org/10.1109/SPMB52430.2021.9672266.

Shin, E., Jang, J. and Baek, S., 'Impacts of Symbolic Regression-Integrated Physics Classes on Students' Perceptions of Using Machine Learning in Science', in: *2024 21st International Conference on Information Technology Based Higher Education and Training (ITHET)*, 2024, pp. 1–8, https://doi.org/10.1109/ITHET61869.2024.10837653.

Si, C., Yang, D. and Hashimoto, T., 'Can LLMs Generate Novel Research Ideas? A Large-Scale Human Study with 100+ NLP Researchers', in: *The Thirteenth International Conference on Learning Representations*, 2024.

Sillitoe, I., Bordin, N., Dawson, N., Waman, V.P., Ashford, P. et al., 'CATH: Increased Structural Coverage of Functional Space', *Nucleic Acids Research*, No. 49, 2021, pp. 266–273, https://doi.org/10.1093/nar/gkaa1079.

Singhi, N., Bansal, H., Hosseini, A., Grover, A., Chang, K.-W. et al., 'When To Solve, When To Verify: Compute-Optimal Problem Solving and Generative Verification for LLM Reasoning', arXiv, 2025, https://doi.org/10.48550/arXiv.2504.01005.

Siozos, P., Hausmann, N., Holst, M. and Anglos, D., 'Application of Laser-Induced Breakdown Spectroscopy and Neural Networks on Archaeological Human Bones for the Discrimination of Distinct Individuals', *Journal of Archaeological Science: Reports*, No. 35, 2021, https://doi.org/10.1016/j.jasrep.2020.102769.

Sommerschield, T., Assael, Y., Pavlopoulos, J., Stefanak, V., Senior, A. et al., 'Machine Learning for Ancient Languages: A Survey', *Computational Linguistics*, No. 49, 2023, pp. 703–747, https://doi.org/10.1162/coli_a_00481.

Steinegger, M., Mirdita, M. and Söding, J., 'Protein-Level Assembly Increases Protein Sequence Recovery from Metagenomic Samples Manyfold', *Nature Methods*, No. 16, Nature Publishing Group, 2019, pp. 603–606, https://doi.org/10.1038/s41592-019-0437-4.

Steinegger, M. and Söding, J., 'Clustering Huge Protein Sequence Sets in Linear Time', *Nature Communications*, No. 9, Nature Publishing Group, 2018, p. 2542, https://doi.org/10.1038/s41467-018-04964-5.

Stephen Cusack, Sebastian Eustermann, Gerard Kleywegt, Jan Kosinski, Julia Mahamid et al., 'Great Expectations – the Potential Impacts of AlphaFold DB', *EMBL Communications*, 2021, https://www.embl.org/news/science/alphafold-potential-impacts.

Stevenson, E., Walsh, C. and Hibberd, L., 'Can Artificial Intelligence Replace Biochemists? A Study Comparing Interpretation of Thyroid Function Test Results by ChatGPT and Google Bard to Practising Biochemists', *Annals of Clinical Biochemistry*, No. 61, SAGE Publications, 2024, pp. 143–149, https://doi.org/10.1177/00045632231203473.

Stocker, M., Oelen, A., Jaradeh, M.Y., Haris, M., Oghli, O.A. et al., 'FAIR Scientific Information with the Open Research Knowledge Graph', *FAIR Connect*, No. 1, SAGE Publications, 2023, pp. 19–21, https://doi.org/10.3233/FC-221513.

Stokes, J.M., Yang, K., Swanson, K., Jin, W., Cubillos-Ruiz, A. et al., 'A Deep Learning Approach to Antibiotic Discovery', *Cell*, No. 181, Elsevier BV, 2020, pp. 475–483, https://doi.org/10.1016/j.cell.2020.04.001.

Straw, I., 'The Automation of Bias in Medical Artificial Intelligence (AI): Decoding the Past to Create a Better Future', *Artificial Intelligence in Medicine*, No. 110, 2020, https://doi.org/10.1016/j.artmed.2020.101965.

Sun, S., Brown, K. and Kusne, A.G., 'Teaching Machine Learning to Materials Scientists: Lessons from Hosting Tutorials and Competitions', *Matter*, No. 5, Elsevier, 2022, pp. 1620–1622, https://doi.org/10.1016/j.matt.2022.04.019.

Swanson, D.R., 'Undiscovered Public Knowledge', *The Library Quarterly*, No. 56, The University of Chicago Press, 1986, pp. 103–118, https://doi.org/10.1086/601720.

Szymanski, N.J., Rendy, B., Fei, Y., Kumar, R.E., He, T. et al., 'An Autonomous Laboratory for the Accelerated Synthesis of Novel Materials', *Nature*, No. 624, 2023, pp. 86–91, https://doi.org/10.1038/s41586-023-06734-w.

Talirz, L., Kumbhar, S., Passaro, E., Yakutovich, A.V., Granata, V. et al., 'Materials Cloud, a Platform for Open Computational Science', *Scientific Data*, No. 7, Nature Publishing Group, 2020, p. 299, https://doi.org/10.1038/s41597-020-00637-5.

Tan, Z., Li, D., Wang, S., Beigi, A., Jiang, B. et al., 'Large Language Models for Data Annotation and Synthesis: A Survey', in: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 930–957, https://doi.org/10.18653/v1/2024.emnlp-main.54.

Tang, X., Zhu, K., Yuan, T., Zhang, Y., Zhou, W. et al., 'Risks of AI Scientists: Prioritizing Safeguarding Over Autonomy', in: *Data in Generative Models - The Bad, the Ugly, and the Greats*, 2025.

The BS Detector, 'AI, Materials, and Fraud, Oh My!', *The BS Detector*, 2025, https://thebsdetector.substack.com/p/ai-materials-and-fraud-oh-my.

The Royal Society, 'Science in the Age of AI - How Artificial Intelligence Is Changing the Nature and Method of Scientific Research',2024.

The UniProt Consortium, 'UniProt: The Universal Protein Knowledgebase in 2025', *Nucleic Acids Research*, No. 53, 2025, pp. 609–617, https://doi.org/10.1093/nar/gkae1010.

Tom, G., Schmid, S.P., Baird, S.G., Cao, Y., Darvish, K. et al., 'Self-Driving Laboratories for Chemistry and Materials Science', *Chemical Reviews*, No. 124, American Chemical Society, 2024, pp. 9633–9732, https://doi.org/10.1021/acs.chemrev.4c00055.

Toner-Rodgers, A., 'Artificial Intelligence, Scientific Discovery, and Product Innovation', *arXiv preprint arXiv:2412.17866*, 2024.

Treloar, N.J., Braniff, N., Ingalls, B. and Barnes, C.P., 'Deep Reinforcement Learning for Optimal Experimental Design in Biology', *PLOS Computational Biology*, No. 18, Public Library of Science, 2022, https://doi.org/10.1371/journal.pcbi.1010695.

Tsamados, A., Floridi, L. and Taddeo, M., 'Human Control of AI Systems: From Supervision to Teaming', *AI and Ethics*, No. 5, 2025, pp. 1535–1548, https://doi.org/10.1007/s43681-024-00489-4.

Tsitsa, I., Conev, A., David, A., Islam, S.A. and Sternberg, M.J.E., 'The AlphaFold Database Ages', bioRxiv, 2025, https://doi.org/10.1101/2025.06.22.660930.

Tyser, K., Segev, B., Longhitano, G., Zhang, X.-Y., Meeks, Z. et al., 'AI-Driven Review Systems: Evaluating LLMs in Scalable and Bias-Aware Academic Reviews', arXiv, 2024, https://doi.org/10.48550/arXiv.2408.10365.

UNESCO, 'Recommendation on the Ethics of Artificial Intelligence' (No. SHS/BIO/PI/2021/1), United Nations Educational, Scientific and Cultural Organization (UNESCO), 2022.

Valizadegan, H., Martinho, M.J.S., Wilkens, L.S., Jenkins, J.M., Smith, J.C. et al., 'ExoMiner: A Highly Accurate and Explainable Deep Learning Classifier That Validates 301 New Exoplanets', *The Astrophysical Journal*, No. 926, The American Astronomical Society, 2022, p. 120, https://doi.org/10.3847/1538-4357/ac4399.

Van, M.-H., Verma, P., Zhao, C. and Wu, X., 'A Survey of AI for Materials Science: Foundation Models, LLM Agents, Datasets, and Tools', arXiv, 2025, https://doi.org/10.48550/arXiv.2506.20743.

Van Noorden, R. and Perkel, J.M., 'AI and Science: What 1,600 Researchers Think', *Nature*, No. 621, 2023.

Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C. et al., 'AlphaFold Protein Structure Database: Massively Expanding the Structural Coverage of Protein-Sequence Space with High-Accuracy Models', *Nucleic Acids Research*, No. 50, 2022a, pp. 439–444, https://doi.org/10.1093/nar/gkab1061.

Varadi, M., Nair, S., Sillitoe, I., Tauriello, G., Anyango, S. et al., '3D-Beacons: Decreasing the Gap between Protein Sequences and Structures through a Federated Network of Protein Structure Data Resources', *GigaScience*, No. 11, 2022b, https://doi.org/10.1093/gigascience/giac118.

Varadi, M. and Velankar, S., 'The Impact of AlphaFold Protein Structure Database on the Fields of Life Sciences', *PROTEOMICS*, No. 23, 2023, https://doi.org/10.1002/pmic.202200128.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L. et al., 'Attention Is All You Need', arXiv, 2023, https://doi.org/10.48550/arXiv.1706.03762.

Vecchietti, L.F., Lee, M., Hangeldiyev, B., Jung, H., Park, H. et al., 'Recent Advances in Interpretable Machine Learning Using Structure-Based Protein Representations', arXiv, 2024, https://doi.org/10.48550/arXiv.2409.17726.

Verschoof-van der Vaart, W.B. and Landauer, J., 'Using CarcassonNet to Automatically Detect and Trace Hollow Roads in LiDAR Data from the Netherlands', *Journal of Cultural Heritage*, No. 47, 2021, pp. 143–154, https://doi.org/10.1016/j.culher.2020.10.009.

Vieriu, A.M. and Petrea, G., 'The Impact of Artificial Intelligence (AI) on Students' Academic Development', *Education Sciences*, No. 15, Multidisciplinary Digital Publishing Institute, 2025, p. 343, https://doi.org/10.3390/educsci15030343.

Wachinger, J., Bärnighausen, K., Schäfer, L.N., Scott, K. and McMahon, S.A., 'Prompts, Pearls, Imperfections: Comparing ChatGPT and a Human Researcher in Qualitative Data Analysis', *Qualitative Health Research*, No. 35, SAGE Publications Inc, 2024, pp. 951–966, https://doi.org/10.1177/10497323241244669.

Wagner, C., Strohmaier, M., Olteanu, A., Kıcıman, E., Contractor, N. et al., 'Measuring Algorithmically Infused Societies', *Nature*, No. 595, 2021, pp. 197–204, https://doi.org/10.1038/s41586-021-03666-1.

Wallwork, A., 'Using AI as an Aid for Writing Papers', in: Wallwork, A. (Ed.), *English for Academic Research: Writing Exercises*, Springer Nature Switzerland, Cham, 2024, pp. 187–212, https://doi.org/10.1007/978-3-031-53174-3_11.

Walmsley, J., 'Artificial Intelligence and the Value of Transparency', *AI & SOCIETY*, No. 36, 2021, pp. 585–595, https://doi.org/10.1007/s00146-020-01066-z.

Walter, Y., 'Embracing the Future of Artificial Intelligence in the Classroom: The Relevance of AI Literacy, Prompt Engineering, and Critical Thinking in Modern Education', *International Journal of Educational Technology in Higher Education*, No. 21, 2024, p. 15, https://doi.org/10.1186/s41239-024-00448-3.

Walters, W.H. and Wilder, E.I., 'Fabrication and Errors in the Bibliographic Citations Generated by ChatGPT', *Scientific Reports*, No. 13, Nature Publishing Group, 2023, p. 14045, https://doi.org/10.1038/s41598-023-41032-5.

Wang, C. and Han, J., 'DL4SciVis: A State-of-the-Art Survey on Deep Learning for Scientific Visualization', *IEEE Transactions on Visualization and Computer Graphics*, No. 29, 2023, pp. 3714–3733, https://doi.org/10.1109/TVCG.2022.3167896.

Wang, G., Frederick, R., Duan, J., Wong, W.B.L., Rupar, V. et al., 'Detecting Misinformation through Framing Theory: The Frame Element-Based Model', *Journal of Computational Social Science*, No. 8, 2025, p. 72, https://doi.org/10.1007/s42001-025-00403-w.

Wang, H., Fu, T., Du, Y., Gao, W., Huang, K. et al., 'Scientific Discovery in the Age of Artificial Intelligence', *Nature*, No. 620, 2023, pp. 47–60, https://doi.org/10.1038/s41586-023-06221-2.

Wang, Jiyao, Hu, H., Wang, Z., Yan, S., Sheng, Y. et al., 'Evaluating Large Language Models on Academic Literature Understanding and Review: An Empirical Study among Early-Stage Scholars', in: *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, Association for Computing Machinery, New York, NY, USA, 2024, pp. 1–18, https://doi.org/10.1145/3613904.3641917.

Wang, Jue, Watson, J.L. and Lisanza, S.L., 'Protein Design Using Structure-Prediction Networks: AlphaFold and RoseTTAFold as Protein Structure Foundation Models', *Cold Spring Harbor Perspectives in Biology*, No. 16, 2024, https://doi.org/10.1101/cshperspect.a041472.

Wang, J. and Zhang, Y., 'Artificial Intelligence in High-Entropy Materials', *Next Materials*, No. 9, 2025, https://doi.org/10.1016/j.nxmate.2025.100993.

Wang, M., Zhang, Z., Bedi, A.S., Velasquez, A., Guerra, S. et al., 'A Call for Built-in Biosecurity Safeguards for Generative AI Tools', *Nature Biotechnology*, No. 43, Nature Publishing Group, 2025, pp. 845–847, https://doi.org/10.1038/s41587-025-02650-8.

Wang, X., Xu, X., Tong, W., Liu, Q. and Liu, Z., 'DeepCausality: A General AI-Powered Causal Inference Framework for Free Text: A Case Study of LiverTox', *Frontiers in Artificial Intelligence*, No. 5, Frontiers, 2022, https://doi.org/10.3389/frai.2022.999289.

Wang, X., Zhang, T., Liu, G., Cui, Z., Zeng, Z. et al., 'LightRoseTTA: High-Efficient and Accurate Protein Structure Prediction Using a Light-Weight Deep Graph Model', *Advanced Science*, No. 12, 2025, https://doi.org/10.1002/advs.202309051.

Wang, Z., Li, Y. and Li, H., 'Chinese Inscription Restoration Based on Artificial Intelligent Models', *npj Heritage Science*, No. 13, Nature Publishing Group, 2025, p. 326, https://doi.org/10.1038/s40494-025-01900-x.

Wasim, T. and Zaheer, A., 'Artificial Intelligence (AI) Revolution in Research: Transforming Data into Discovery', *Esculapio*, No. 19, 2023, pp. 376–377, https://doi.org/10.51273/esc23.251319426.

Wätzold, F., Popiela, B. and Mayer, J., 'Methodology for AI-Based Search Strategy of Scientific Papers: Exemplary Search for Hybrid and Battery Electric Vehicles in the Semantic Scholar Database', *Publications*, No. 12, 2024, p. 49, https://doi.org/10.3390/publications12040049.

Webb, C.A., Ren, B., Rahimi-Eichi, H., Gillis, B.W., Chung, Y. et al., 'Personalized Prediction of Negative Affect in Individuals with Serious Mental Illness Followed Using Long-Term Multimodal Mobile Phenotyping', *Translational Psychiatry*, No. 15, Nature Publishing Group, 2025, p. 174, https://doi.org/10.1038/s41398-025-03394-4.

Weber-Wulff, D., Anohina-Naumeca, A., Bjelobaba, S., Foltýnek, T., Guerrero-Dib, J. et al., 'Testing of Detection Tools for AI-Generated Text', *International Journal for Educational Integrity*, No. 19, 2023, p. 26, https://doi.org/10.1007/s40979-023-00146-z.

Wei, H. and McCammon, J.A., 'Structure and Dynamics in Drug Discovery', *npj Drug Discovery*, No. 1, Nature Publishing Group, 2024, pp. 1–8, https://doi.org/10.1038/s44386-024-00001-2.

Weiskopf, D., 'Bridging Quantitative and Qualitative Methods for Visualization Research: A Data/Semantics Perspective in Light of Advanced AI', 2024.

Weissenow, K., Heinzinger, M., Steinegger, M. and Rost, B., 'Ultra-Fast Protein Structure Prediction to Capture Effects of Sequence Variation in Mutation Movies', bioRxiv, 2022, https://doi.org/10.1101/2022.11.14.516473.

Wenger, E., 'Communities of practice: Learning, meaning, and identity', Cambridge University Press, 1998.

Wenger, E. and Kenett, Y., 'We're Different, We're the Same: Creative Homogeneity Across LLMs', arXiv, 2025, https://doi.org/10.48550/arXiv.2501.19361.

Whitfield, S. and Hofmann, M.A., 'Elicit: AI Literature Review Research Assistant', *Public Services Quarterly*, No. 19, Routledge, 2023, pp. 201–207, https://doi.org/10.1080/15228959.2023.2224125.

Wilkinson, M.D., Dumontier, M., Aalbersberg, Ij.J., Appleton, G., Axton, M. et al., 'The FAIR Guiding Principles for Scientific Data Management and Stewardship', *Scientific Data*, No. 3, 2016, https://doi.org/10.1038/sdata.2016.18.

Wills, S., Poon, S.T.S., Salili-James, A. and Scott, B., 'The Use of Generative AI for Coding in Academia', *Methods in Ecology and Evolution*, No. 15, 2024, pp. 2189–2191, https://doi.org/10.1111/2041-210X.14454.

Wohlwend, J., Corso, G., Passaro, S., Reveiz, M., Leidal, K. et al., 'Boltz-1 Democratizing Biomolecular Interaction Modeling', *bioRxiv: The Preprint Server for Biology*, 2024, https://doi.org/10.1101/2024.11.19.624167.

Wong, F., Krishnan, A., Zheng, E.J., Stärk, H., Manson, A.L. et al., 'Benchmarking AlphaFold-enabled Molecular Docking Predictions for Antibiotic Discovery', *Molecular Systems Biology*, No. 18, John Wiley & Sons, Ltd, 2022, https://doi.org/10.15252/msb.202211081.

Wu, A., Wang, Y., Shu, X., Moritz, D., Cui, W. et al., 'AI4VIS: Survey on Artificial Intelligence Approaches for Data Visualization', *IEEE Transactions on Visualization and Computer Graphics*, No. 28, 2022, pp. 5049–5070, https://doi.org/10.1109/TVCG.2021.3099002.

Wu, H., Wang, W., Wan, Y., Jiao, W. and Lyu, M., 'ChatGPT or Grammarly? Evaluating ChatGPT on Grammatical Error Correction Benchmark', arXiv, 2023, https://doi.org/10.48550/arXiv.2303.13648.

Wu, R., Ding, F., Wang, R., Shen, R., Zhang, X. et al., 'High-Resolution de Novo Structure Prediction from Primary Sequence', bioRxiv, 2022, https://doi.org/10.1101/2022.07.21.500999.

Wu, S., Irsoy, O., Lu, S., Dabravolski, V., Dredze, M. et al., 'BloombergGPT: A Large Language Model for Finance', arXiv, 2023, https://doi.org/10.48550/arXiv.2303.17564.

Wu, Y., Wan, Y., Zhang, H., Sui, Y., Wei, W. et al., 'Automated Data Visualization from Natural Language via Large Language Models: An Exploratory Study', *Proceedings of the ACM on Management of Data*, No. 2, 2024, pp. 1–28, https://doi.org/10.1145/3654992.

Xie, T., Fu, X., Ganea, O.-E., Barzilay, R. and Jaakkola, T.S., 'Crystal Diffusion Variational Autoencoder for Periodic Material Generation', in: *The Tenth International Conference on Learning Representations*, 2021.

Xie, Y., Pan, Y., Xu, H. and Mei, Q., 'Bridging AI and Science: Implications from a Large-Scale Literature Analysis of AI4Science', 2024.

Xu, Jian, Yu, C., Xu, Jiawei, Torvik, V.I., Kang, J. et al., 'PubMed Knowledge Graph 2.0: Connecting Papers, Patents, and Clinical Trials in Biomedical Science', *Scientific Data*, No. 12, 2025, https://doi.org/10.1038/s41597-025-05343-8.

Xu, R., Sun, Y., Ren, M., Guo, S., Pan, R. et al., 'AI for Social Science and Social Science of AI: A Survey', *Information Processing & Management*, No. 61, 2024, https://doi.org/10.1016/j.ipm.2024.103665.

Yacoub, B., Varga-Szemes, A., Schoepf, U.J., Kabakus, I.M., Baruah, D. et al., 'Impact of Artificial Intelligence Assistance on Chest CT Interpretation Times: A Prospective Randomized Study', *American Journal of Roentgenology*, No. 219, American Roentgen Ray Society, 2022, pp. 743–751, https://doi.org/10.2214/AJR.22.27598.

Yang, C.-L., Uhde, A., Yamashita, N. and Kuzuoka, H., 'Understanding and Supporting Peer Review Using AI-Reframed Positive Summary', in: *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, Association for Computing Machinery, New York, NY, USA, 2025, pp. 1–16, https://doi.org/10.1145/3706598.3713219.

Yang, D.-H., Zhou, X., Wang, X.-Y. and Huang, J.-P., 'Micro-Earthquake Source Depth Detection Using Machine Learning Techniques', *Information Sciences*, No. 544, 2021, pp. 325–342, https://doi.org/10.1016/j.ins.2020.07.045.

Yang, K., Ji, S., Zhang, T., Xie, Q., Kuang, Z. et al., 'Towards Interpretable Mental Health Analysis with Large Language Models', in: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023, pp. 6056–6077, https://doi.org/10.18653/v1/2023.emnlp-main.370.

Yang, S., Cho, K., Merchant, A., Abbeel, P., Schuurmans, D. et al., 'Scalable Diffusion for Materials Generation', in: *NeurIPS 2023 AI for Science Workshop*, 2023.

Ye, A., Moore, J., Novick, R. and Zhang, A.X., 'Language Models as Critical Thinking Tools: A Case Study of Philosophers', arXiv, 2024, https://doi.org/10.48550/arXiv.2404.04516.

Zeng, Q., Sidhu, M., Blume, A., Chan, H.P., Wang, L. et al., 'Scientific Opinion Summarization: Paper Meta-Review Generation Dataset, Methods, and Evaluation', in: Yin, W., Ahn, J.J., Zhang, R., Huang, L., Hadfi, R. et al. (Eds.), *Artificial Intelligence for Research and Democracy*, Springer Nature, Singapore, 2025, pp. 20–38, https://doi.org/10.1007/978-981-97-9536-9_2.

Zeni, C., Pinsler, R., Zügner, D., Fowler, A., Horton, M. et al., 'A Generative Model for Inorganic Materials Design', *Nature*, Nature Publishing Group UK London, 2025, pp. 1–3.

Zhai, C., Wibowo, S. and Li, L.D., 'The Effects of Over-Reliance on AI Dialogue Systems on Students' Cognitive Abilities: A Systematic Review', *Smart Learning Environments*, No. 11, 2024, p. 28, https://doi.org/10.1186/s40561-024-00316-7.

Zhang, B., Anderljung, M., Kahn, L., Dreksler, N., Horowitz, M.C. et al., 'Ethics and Governance of Artificial Intelligence: Evidence from a Survey of Machine Learning Researchers', *Journal of Artificial Intelligence Research*, No. 71, 2021, pp. 591–666, https://doi.org/10.1613/jair.1.12895.

Zhang, Y., Chen, X., Jin, B., Wang, S., Ji, S. et al., 'A Comprehensive Survey of Scientific Large Language Models and Their Applications in Scientific Discovery', in: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 8783–8817, https://doi.org/10.18653/v1/2024.emnlp-main.498.

Zhao, Fengxiang, Yu, F. and Shang, Y., 'A New Method Supporting Qualitative Data Analysis Through Prompt Generation for Inductive Coding', in: *2024 IEEE International Conference on Information Reuse and Integration for Data Science (IRI)*, 2024, pp. 164–169, https://doi.org/10.1109/IRI62200.2024.00043.

Zhao, Fei, Zhang, C. and Geng, B., 'Deep Multimodal Data Fusion', *ACM Comput. Surv.*, No. 56, 2024, pp. 1-36, https://doi.org/10.1145/3649447.

Zhao, Q., Stalin, S., Zhao, C.-Z. and Archer, L.A., 'Designing Solid-State Electrolytes for Safe, Energy-Dense Batteries', *Nature Reviews Materials*, No. 5, Nature Publishing Group, 2020, pp. 229–252, https://doi.org/10.1038/s41578-019-0165-5.

Zhao, Y., Al-Fahdi, M., Hu, M., Siriwardane, E.M.D., Song, Y. et al., 'High-Throughput Discovery of Novel Cubic Crystal Materials Using Deep Generative Neural Networks', *Advanced Science*, No. 8, 2021, https://doi.org/10.1002/advs.202100566.

Zhao, Z.-J., Liu, S., Zha, S., Cheng, D., Studt, F. et al., 'Theory-Guided Design of Catalytic Materials Using Scaling Relationships and Reactivity Descriptors', *Nature Reviews Materials*, No. 4, Nature

Publishing Group, 2019, pp. 792–804, https://doi.org/10.1038/s41578-019-0152-x.

Zhu, F., Nowaczynski, A., Li, R., Xin, J., Song, Y. et al., 'ScaleFold: Reducing AlphaFold Initial Training Time to 10 Hours', arXiv, 2024, https://doi.org/10.48550/arXiv.2404.11068.

Zhu, J., Zhao, X., Sun, Y., Song, S. and Yuan, X., 'Relational Data Cleaning Meets Artificial Intelligence: A Survey', *Data Science and Engineering*, No. 10, 2025, pp. 147–174, https://doi.org/10.1007/s41019-024-00266-7.

Zhu, R., Nong, W., Yamazaki, S. and Hippalgaonkar, K., 'WyCryst: Wyckoff Inorganic Crystal Generator Framework', *Matter*, No. 7, 2024, pp. 3469–3488, https://doi.org/10.1016/j.matt.2024.05.042.

Zhu, S., Xue, H., Nie, N., Zhu, C., Liu, H. et al., 'Reproducing the Past: A Dataset for Benchmarking Inscription Restoration', in: *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, Association for Computing Machinery, New York, NY, USA, 2024, pp. 7714–7723, https://doi.org/10.1145/3664647.3680587.

Zhuang, T. and Lin, Z., 'The Why, What, and How of AI-Based Coding in Scientific Research', arXiv, 2024, https://doi.org/10.48550/arXiv.2410.02156.

Zhuang, Z., Chen, J., Xu, H., Jiang, Y. and Lin, J., 'Large Language Models for Automated Scholarly Paper Review: A Survey', arXiv, 2025, https://doi.org/10.48550/arXiv.2501.10326.

Ziems, C., Held, W., Shaikh, O., Chen, J., Zhang, Z. et al., 'Can Large Language Models Transform Computational Social Science?', *Computational Linguistics*, No. 50, 2024, pp. 237–291, https://doi.org/10.1162/coli_a_00502.

Zubatiuk, T. and Isayev, O., 'Development of Multimodal Machine Learning Potentials: Toward a Physics-Aware Artificial Intelligence', *Accounts of Chemical Research*, No. 54, American Chemical Society, 2021, pp. 1575–1585, https://doi.org/10.1021/acs.accounts.0c00868.

LIST OF ABBREVIATIONS AND DEFINITIONS

Abbreviations	Definitions
2D	Two-dimensional
3D	Three-dimensional
ADS	Archaeology Data Service
AI	Artificial Intelligence
AIO	Asia Inscriptions Online
ALLEA	All European Academies
AR	Augmented Reality
ARIADNE	Advanced Research Infrastructure for Archaeological Dataset Networking in Europe
ASR-net	Ancient Script Recognition Network
BCE	Before the Common Era
BFD	Big Fantastic Database
BMRB	Biological Magnetic Resonance Data Bank
CARE	Collective benefit, Authority to control, Responsibility and Ethics
CASP	Critical Assessment of Structure Prediction
CDSE	Copernicus Data Space Ecosystem
CERIT	Centre for Education, Research and Innovation in Information and Communication Technologies
CA	Canada
CH	Switzerland
CIF	Crystallographic Information Files
CIRI	Chinese Inscription Rubbing Image
CN	China
CNN	Convolutional Neural Network
CPU	Central Processing Unit
CV	Computer Vision
DANS	Data Archiving and Networked Services
DARIAH	Digital Research Infrastructure for the Arts and Humanities

DFT	Density Functional Theory
DiaDEM	Digital Discovery platform for Organic Electronic Materials
DGTES	Digital Techno-Economic Ecosystem
DL	Deep Learning
DNA	Deoxyribonucleic Acid
DNN	Deep Neural Network
EAGLE	Electronic Archive of Greek and Latin Epigraphy
EBI	European Bioinformatics Institute
EC	European Commission
EDAK	Epigraphische Datenbank zum Antiken Kleinasien (*EN*: Epigraphic Database of Ancient Asia Minor)
EDR	Epigraphic Database Roma
EIC	European Innovation Council
EMBL	European Molecular Biology Laboratory
EMDB	Electron Microscopy Data Bank
EpHEMERA	Online 3D Database System for Endangered architectural and archaeological Heritage in the south Eastern MEditerRAnean area
EPO	European Patent Office
ERA	European Research Area
ESDIS	Earth Science Data and Information System
EU	European Union
EU27	European Union's 27 member states
FAIR	Findable, Accessible, Interoperable and Reusable
FLOP	Floating-Point Operation
GAN	Generative Adversarial Network
GDPR	General Data Protection Regulation
GenAI	Generative Artificial Intelligence
GIS	Geographic Information System
GNN	Graph Neural Network
GPAI	General-Purpose Artificial Intelligence
GPU	Graphics Processing Unit
GUI	Graphical User Interface
HPC	High-Performance Computing
IN	India
IVC	Indus Valley Civilisation

JRC	Joint Research Centre
KR	South Korea
LDB	Literature-Based Discovery
LIBS	Laser-Induced Breakdown Spectroscopy
LiDAR	Light Detection and Ranging
LLM	Large Language Model
LSTM	Long-Short Term Memory
MAST-ML	Materials Simulation Toolkit for Machine Learning
MI-net	Motif Identification Network
ML	Machine Learning
MSA	Multiple Sequence Alignment
MSCA	Marie Skłodowska-Curie Actions
MSRM	Multi-Scale Relief Model
NASA	National Aeronautics and Space Administration
NLP	Natural Language Processing
NMR	Nuclear Magnetic Resonance
NN	Neural Network
NOMAD	Novel Materials Discovery
OCIANA	The Online Corpus of the Inscriptions from Ancient North Arabia
OECD	Organisation for Economic Co-operation and Development
OQMD	Open Quantum Materials Database
PDB	Protein Data Bank
PDBe	Protein Data Bank in Europe
PDBj	Protein Data Bank Japan
PIR	Protein Information Resource
pLDDT	predicted Local Distance Difference Test
RCSB PDB	Research Collaboratory for Structural Bioinformatics Protein Data Bank
REMI	REsource for Materials Informatics
RePAIR	Reconstructing the Past: Artificial Intelligence and Robotics meet Cultural Heritage
RF	Random Forest
RNA	Ribonucleic Acid
RoW	Rest of the world
RS	Recommender/Recommendation System
SDI	Spatial Data Infrastructure

SG	Singapore
SIB	Swiss Institute of Bioinformatics
STS	Science and Technology Studies
SVM	Support Vector Machine
tDAR	the Digital Archaeological Record
TPU	Tensor Processing Unit
UI	User Interface
UK	United Kingdom
UNESCO	United Nations Educational, Scientific and Cultural Organization
US	United States (of America)
USGS	United States Geological Survey
UX	User Experience
VAE	Variational Autoencoder
VC	Venture Capital
VR	Virtual Reality
WHO	World Health Organization
wwPDB	Worldwide Protein Data Bank

LIST OF FIGURES

LIST OF TABLES

Getting in touch with the EU

In person

All over the European Union there are hundreds of Europe Direct centres. You can find the address of the centre nearest you online (european-union.europa.eu/contact-eu/meet-us_en).

On the phone or in writing

Europe Direct is a service that answers your questions about the European Union. You can contact this service:

— by freephone: 00 800 6 7 8 9 10 11 (certain operators may charge for these calls),
— at the following standard number: +32 22999696,
— via the following form: european-union.europa.eu/contact-eu/write-us_en.

Finding information about the EU

Online

Information about the European Union in all the official languages of the EU is available on the Europa website (european-union.europa.eu).

EU publications

You can view or order EU publications at op.europa.eu/en/publications. Multiple copies of free publications can be obtained by contacting Europe Direct or your local documentation centre (european-union.europa.eu/contact-eu/meet-us_en).

EU law and related documents

For access to legal information from the EU, including all EU law since 1951 in all the official language versions, go to EUR-Lex (eur-lex.europa.eu).

Open data from the EU

The portal data.europa.eu provides access to open datasets from the EU institutions, bodies and agencies. These can be downloaded and reused for free, for both commercial and non-commercial purposes. The portal also provides access to a wealth of datasets from European countries.

Science for policy

The Joint Research Centre (JRC) provides independent, evidence-based knowledge and science, supporting EU policies to positively impact society

Scan the QR code to visit:

The Joint Research Centre: EU Science Hub

https://joint-research-centre.ec.europa.eu

Publications Office
of the European Union